
# Enhancement in Meta Clustering Using Probability Distribution to Improve the Cluster Quality

A Dissertation

Submitted

**By**

**Veerpal**

**11303463**

To

**Department of Computer Science and Engineering**

In partial fulfillment of the Requirement for the

Award of the Degree of

**Master of Technology in Computer Science and Engineering**

**Under the guidance of**

**Mr. Sanyam Anand**

**(May 2015)**

# PAC APPROVAL

INDIA'S LARGEST UNIVERSITY

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

School of: Science & Technology

**DISSERTATION TOPIC APPROVAL PERFORMA**

Name of the Student: Veerpal

Registration No: 11303463

Batch: 2013-2015

Roll No. B42

Session: 2014-2015

Parent Section: RK2305

**Details of Supervisor:**

Designation: A.P

Name: SANYAM

Qualification: M.Tech

U.ID: 15736

Research Experience: 2 yrs

**SPECIALIZATION AREA:** Data mining **(pick from list of provided specialization areas by DAA)**

**PROPOSED TOPICS**

1. Data mining (Data preprocessing (outlier detection & clustering))
2. Association Rules
3. Big data

15736

Signature of Supervisor

**PAC Remarks:**

Topic 1 is approved.

11/11 21/9/14

**APPROVAL OF PAC CHAIRPERSON:** **Signature:** **Date:**

*Supervisor should finally encircle one topic out of three proposed topics and put up for approval before Project Approval Committee (PAC)

*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.

*One copy to be submitted to Supervisor.

## ABSTRACT

Data mining is taking out of hidden patterns from huge amount of data. Clustering is an analyzing of cluster basically the old-fashioned subjects in the data mining arena. Clustering can be categorized into various types. Meta clustering is one of the clustering types. Digital data is mostly uncertain in nature and handling that digital data is quite difficult. In this paper, an enhancement in Meta clustering has been proposed to improve the accuracy in the clusters.

## CERTIFICATE

This is to certify that Veerpal has completed M.Tech dissertation titled **Enhancement in Meta clustering using probability distribution to improve the cluster quality** under my guidance and supervision. To the best of my knowledge, present work is the result of her original investigation and study. No part of the dissertation has ever been submitted for any other degree or diploma. The dissertation is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech Computer Science & Engg.

Date----------

Signature of Advisor

Name: Sanyam Anand

UID: 15736

## ACKNOWLEDGEMENT

First and foremost, I want to thank the Department of CSE of Lovely Professional University for giving me permission to begin Thesis in first instance, to do necessary research work and to use required data. I would like to acknowledge the assistance provided to me by the library staff of L.P.U.

Inspiration to action is the most important ingredient required throughout the task. I am deeply indebted to my mentor **Mr. Sanyam Anand** whose help, stimulating suggestions and encouragement helped me in all the time of research.

I express my gratitude to my parents for being a continuous source of encouragement and for their financial aids given to me. Finally, I would like to express my gratitude to all those who helped and supported me.

## DECLARATION

I hereby declare that the dissertation entitled, **"Enhancement in Meta clustering using probability distribution to improve the cluster quality"** submitted for the M. Tech. Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: **2 May, 2015**

Investigator

**Reg. No. 11303463**

## TABLE OF CONTENTS

## LIST OF FIGURES

**List of Figures** **Page no.**

# CHAPTER-1
# INTRODUCTION

There are wonderful dimensions of data filling networks, computers and lives. Science institutes, government agencies and business organizations collect data from various resources. This collected data is in very large amount. Data collected from various resources is actually not usable. Data is increasing day by day in volume, variety and velocity. Data has basically three features volume i.e. data is very big in size when collected from various resources. Second feature of data is velocity means data is moving in nature. Third feature of data is variety i.e. data is varying in nature, it varies day by day. Actually small amount of data is usable not all data collected is usable. Simply data volumes are very large and are difficult to manage these data becomes complicated structures and becomes more complex. Data mining is used to take only usable data from these volumes of data. Data mining is just because to understand the nature of data and to handle data. In today's competitive world data mining is very important i.e. extract beneficial knowledge hidden in large volumes of data. Computer based methodology and new approaches are used to realize knowledge from large volumes of data which is called as data mining. Automatic and manual methods used to describe the progress of the data mining process. It is an iterative process. Data mining means searching new information from large volumes of data. It contain basically two goals first is prediction and second one is description. Prediction contains variables that predict the missing values of data or unknown values. On the other hand description finds the patterns that are interesting to user. Data mining activities are as follows.

1. **Predictive data mining: -** Given data sets predict the model which is produced by predictive data mining. It produces the model for mining information from huge amount of data.

2. **Descriptive data mining**: - In this nontrivial information is generated based on the given data sets. Descriptive data mining describes about the model that is used to extract the useful knowledge from large data volumes.

Model produced in predictive data mining is in executable code. This code is used to describe classification, prediction, estimation and some other tasks. Data mining primary tasks are as listed below

1. **Classification-** Classification is defined as it describes the predictive learning function. This function divides the data items into one of numerous predefined classes.

2. **Regression-** It also describes the predictive learning function. This function maps the data item to real values prediction variables.

3. **Clustering-** It is a descriptive task in which similar data items are grouped together and dissimilar are discarded as noise. Within a cluster data items are more similar to each other.

4. **Summarization-** It is also a descriptive task it contain the method for finding compact description of set of data.

5. **Dependency modeling-** It describe the local model which describe the dependencies among the variable or between the values of data sets

6. **Change and Deviation Detection**- It describes the maximum substantial variations in data set.

Data mining is most and fast growing field in computer. There are wide range of techniques and methodologies that are used to mine data. Data mining is just like mining coal from rocks. Means extract usable data from large volumes of data sets. Fraudulent activities and crime trends discovered using data mining. It is very useful in data extraction. Data mining has basically two origins statistics and machine learning. Statistic is basically originated from mathematics. Machine learning's origin is computer practice.

**1.2 Data mining process:** Data mining process include number of steps that identify the problem and fix the solution for that particular problem. It include following steps.

1. State the problem and formulate the hypothesis
2. Collect the data

3. Preprocessing the data
4. Estimate the model
5. Interpret the model and draw conclusion

Above five steps are there to identify the problem and fix the solution for that particular problem this can be described diagrammatically as given below.

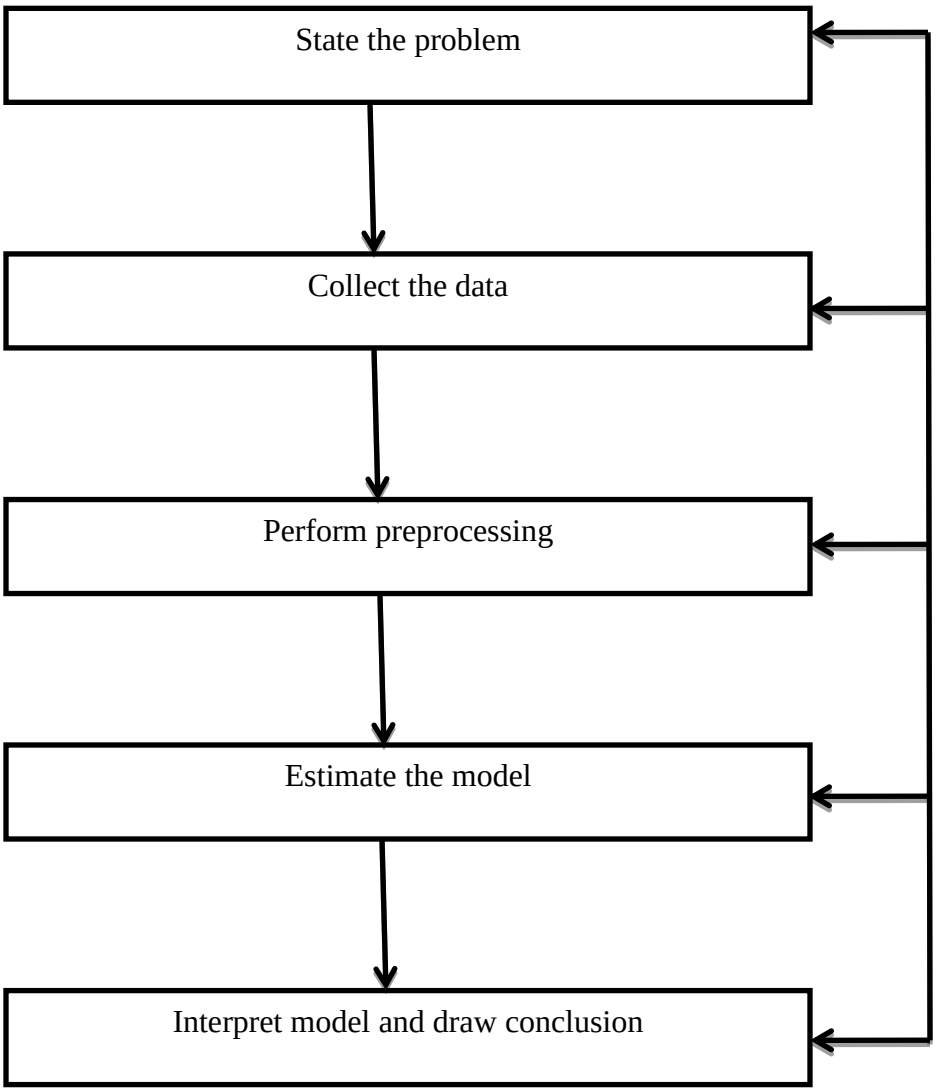


**Fig 1.1 Data mining Process**

1. **State the problem: -** Data mining process include state problem it is the first step of data mining process. It describe about the problem that exist in data and formulate the hypothesis to solve the problem. Hypothesis is assumptions that are required to solve the problem.

2. **Collect the data: -** This step contains how actually data are generated and collected from various resources. This basically includes two steps. Data generation under the control of expert that is known as designed experiment. Second approach is collect data under the observation is called observational approach. Collection of data play important role in data mining process.

3. **Preprocessing the data: -** Data are mostly collected from existing data sources, data marts and data warehouse. There may present missing values in data or say unknown values in data may present, to clean data, data preprocessing is performed. There are two common tasks that are performed in data preprocessing.

   **Outlier detection and removal: -** outliers are values that are not consistent with other data sets. Outliers are detected are recording errors, measurement errors and abnormal values. These can be detected and removed using data preprocessing. By developing a modeling method outliers can be removed.

   **Scaling, encoding and selecting features:** - Data preprocessing contain various steps like encoding and variable scaling. Also dimension reduction is performed in data preprocessing. This step is not independent of other phases of data mining process. It totally depends upon the all another steps described in data mining.

4. **Estimate the model:** - Selecting appropriate data mining technique and implementation of that technique is main task of this phase. There are various models of implementation and selection of one optimal solution model.

5. **Interpret the model and Draw the conclusion: -** Model of data mining process should help in decision making. Model should be easily understandable and decision making because one can't make their decision on the basis of black box model. Finally one can draw conclusion from the model.

Data preprocessing include data collected from data warehouse as given below in diagram. Data warehouse is also a part of data mining. Data mining include data warehouse. Data warehouse comprise historical data in it. Historical information is presented here in data warehouse. Integrated data is presented in data warehouse. These are the central repositories of collected historical data from various resources. Data warehouse is nonvolatile in nature. Data is permanently stored in data warehouse. Online analytical processing of data is presented in data warehouse.

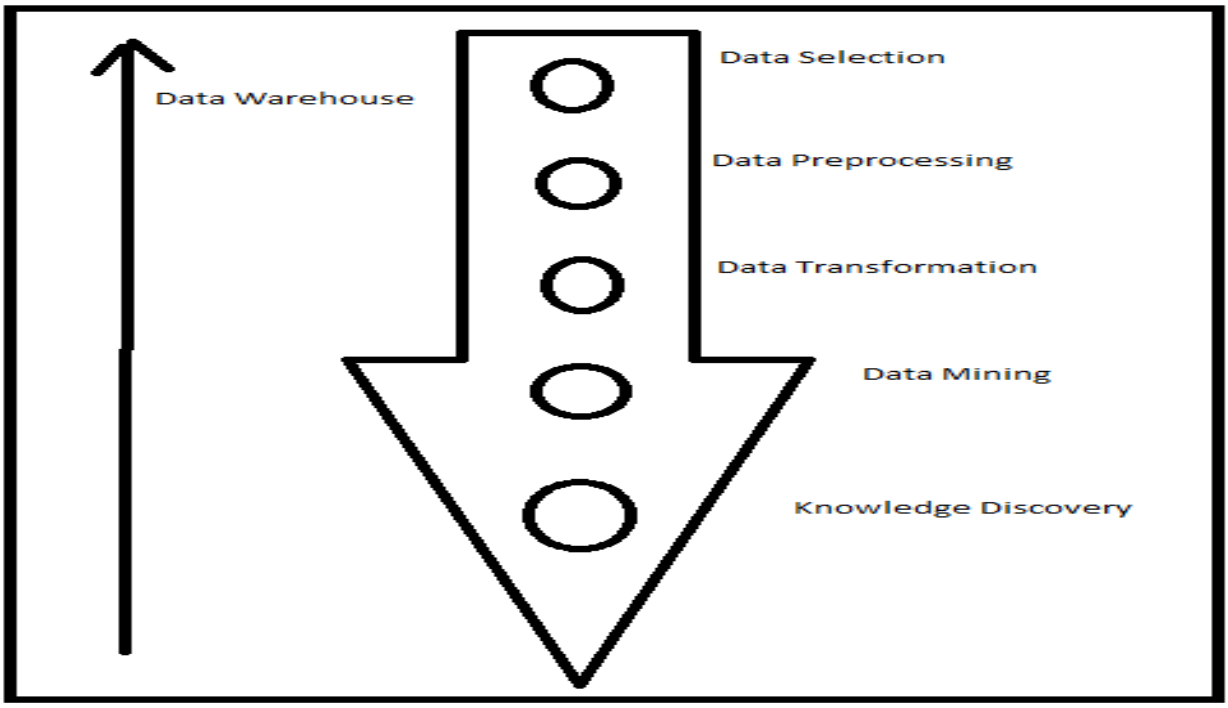


**Fig 1.2 Data warehouse**

**1.3 Clustering in data mining: -** clustering is defined as process of combining similar data items into a single group is called as clustering. Clusters are extracted in the form of groups. Clustering is unsupervised learning that cluster or group natural data items. It is actually a process of grouping abstract into class of similar objects. Clusters are group of similar objects. Clustered data objects are more similar to each other than uncluttered data objects. Density data is present for clustering. Clustering is actually partitioning of data sets into meaningful classes it helps the users to understand the real structure of data sets. There is no predefined classes are there in clustering because it is unsupervised learning. Data structure can be easily found using clustering. Clustering method decide the quality of cluster. Hidden

pattern discovery can measure the cluster quality. Clustering is procedure in which similar data elements are combined together and dissimilar are removed, in clustering different documents are grouped in a single group. In this same documents or say similar documents are grouped in a same cluster. Many advantages are there but alternative advantage of clustering is that document will not misplace. In case a document is misplaced then it can be easily found by using clustering algorithms. Clustering contain many applications areas. These application areas are as listed below.

1. Economic science
2. Document classification
3. Spatial data analysis
4. Pattern recognition
5. Image processing

There are several examples of clustering applications. These applications are marketing, land use, insurance, city-planning and many other applications that are present of clustering are.

**Clustering contain following categories as listed below.**

1. Partitioning algorithm
2. Hierarchy algorithm
3. Density based clustering
4. Grid based clustering
5. Model based clustering

**1. Partitioning clustering algorithm**

Partitioning clustering algorithm produces k number of partition containing n data objects. K number of clusters produces using partitioning clustering method. There are many clusters, all are filled, there is no empty cluster in data partitioning clustering algorithm. N is input parameter and k are the output number of clusters. Partitioning clustering algorithm is basically of two types.

K-mean clustering algorithm

K-Medoid clustering algorithm

K- Median clustering algorithm

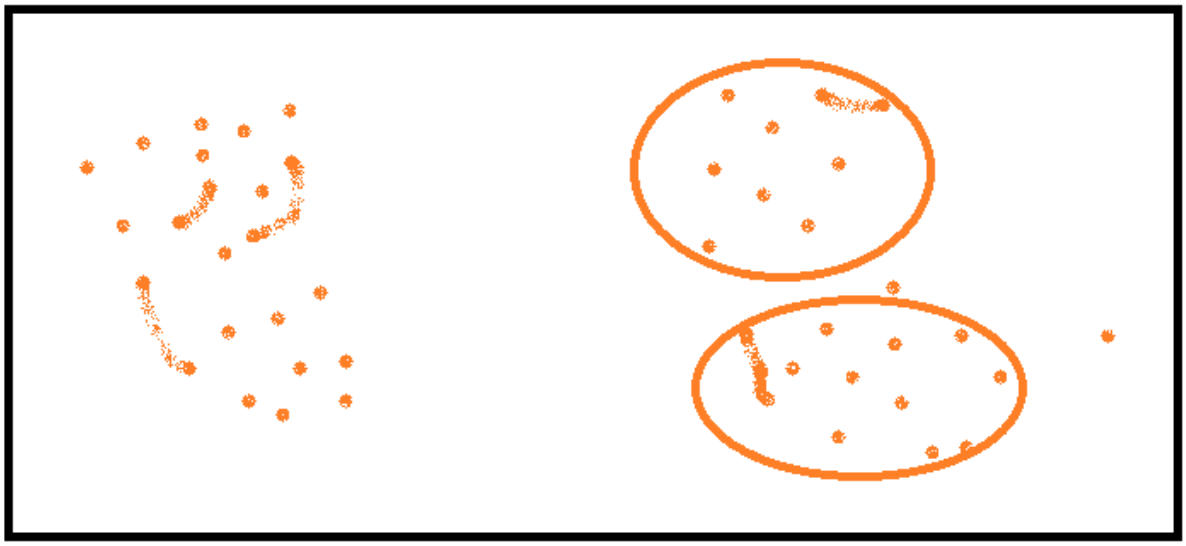

**Fig 1.3 Partitioning clustering algorithm**

**Hierarchal clustering algorithm: -** Hierarchal clustering algorithm follow basically two approaches i.e. bottom up approach and top down approach. Both approaches collectively produce dendrogram. It starts by using single instance clusters. Each and every step includes joining of two clusters. Bottom up approach is also called as agglomerative approach. Top down approach is also called divisive approach or say deglomerative approach. Top down approach starts with one universal cluster. It can be very fast.

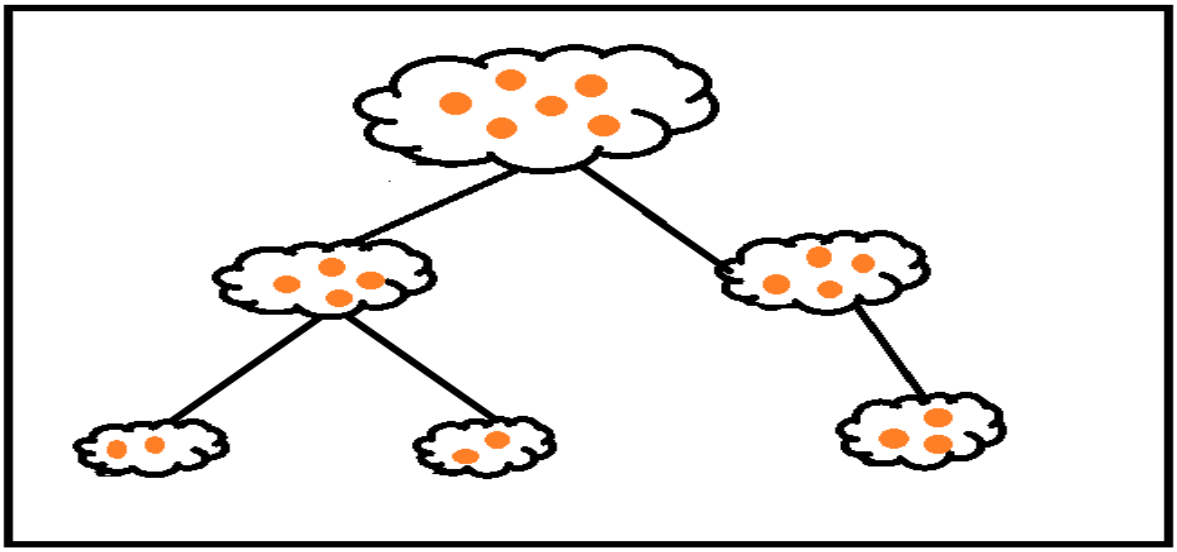

**Fig. 1.4Hierarchical Clustering**

## 2. Density Based Clustering

Density based clustering is defined as clustering based on density local cluster formation such as density associated points. Each cluster has considerable density points within the cluster. Density based clustering include DBSCAN. It handles the noise of data. It requires density parameters and used to define the clusters of arbitrary shape. In density based clustering DBSCAN is basic algorithm firstly a radius parameter is selected and then according to min points data is clustered accordingly on the basis of min point's concept.

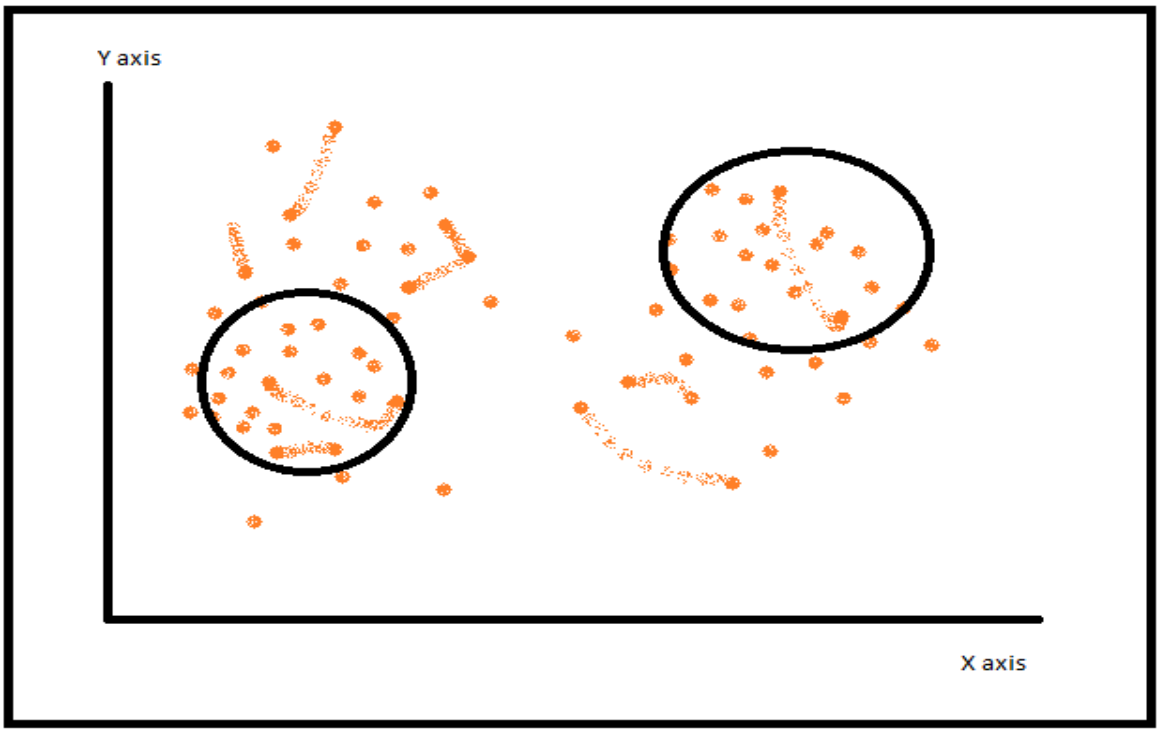


**Fig. 1.5 Density based clustering**

3. **Grid based clustering: -** In this objects are cooperatively assembled together to form network. Item distance is determined into limited number of booths that produce a network assembly. In this first of all compute density of each cell. After that eliminate whose density is below threshold value. Now form cluster according to group of dense clusters. In this no distance computations so it is fast process. In this it is also easy to determine which cluster is neighboring. Here in this figures are narrow to the union. As shown in fig number of different cells are presented over here to form the

grids of different data sets. It forms the grids of similar data item. Grouping of similar data items is given in the form of grids.

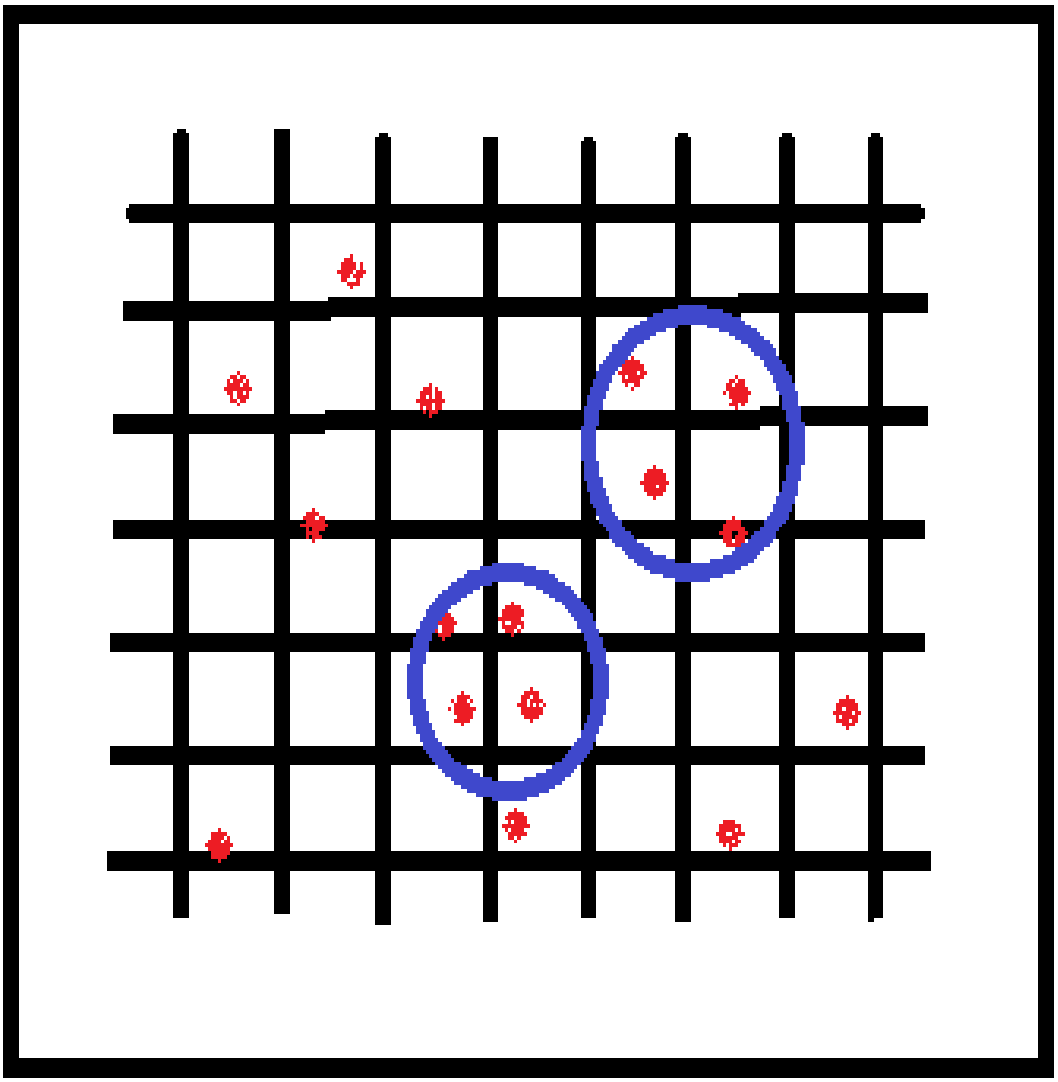

**Fig 1.6 grid based clustering**

4. **Model based clustering: -** Model based clustering is used to fit the data and mathematical model into a solution of clustering. In model based clustering there are basically two approaches one is artificial intelligence and another is statistical. It would describe about the characteristics of each object

### 1.4 Requirements of clustering

1. High dimensionality
2. Able to deal with noise and outliers
3. Arbitrary shaped clusters discovery
4. Usability
5. Interpretability
6. Dealing with different types of attributes
7. To determine input parameters minimal requirements of domain knowledge
8. Scalability

**1.5 Meta clustering: -** Meta clustering is defined as clustering of clustering. It is used to cluster the various clustering together which are similar to each other. Basically three steps are followed that help to perform meta-clustering.

1. Produce some good base level clusterings of same data sets which are quantitatively dissimilar to each other
2. Measure the similarity between base level clusterings that are produced in first step to cluster the base level clusterings.
3. Present the meta-clustering to user by consolidating the base level clusterings.

Grouping of different levels of clustering is present over here in meta-clustering. K-mean is the base level clustering that is applied to perform meta-clustering. After applying base level clustering some other clusterings are applied to remove the limitations of base level clustering. After base level clustering spectral clustering is applied. Adjustment factor is applied at the end to improve the cluster quality. Following steps are there that meta-clustering follow as given below.

1. Arrange data items
2. Apply basic clustering i.e. K-mean clustering
3. Apply spectral clustering to improve the clustering i.e. text clustering and spectral clustering.
4. Apply adjustment factor for better quality.

**1.5.1 Arrange Data:** Data is a raw fact. In data mining it is arrangement of data elements randomly on 2- D plane to give its structure and apply clustering mechanism. Clustering is applied on data sets in two dimensional planes. First of all data scattered in two dimensional plane and then arrange the data in a manner so that it can be easily clustered into clusters.

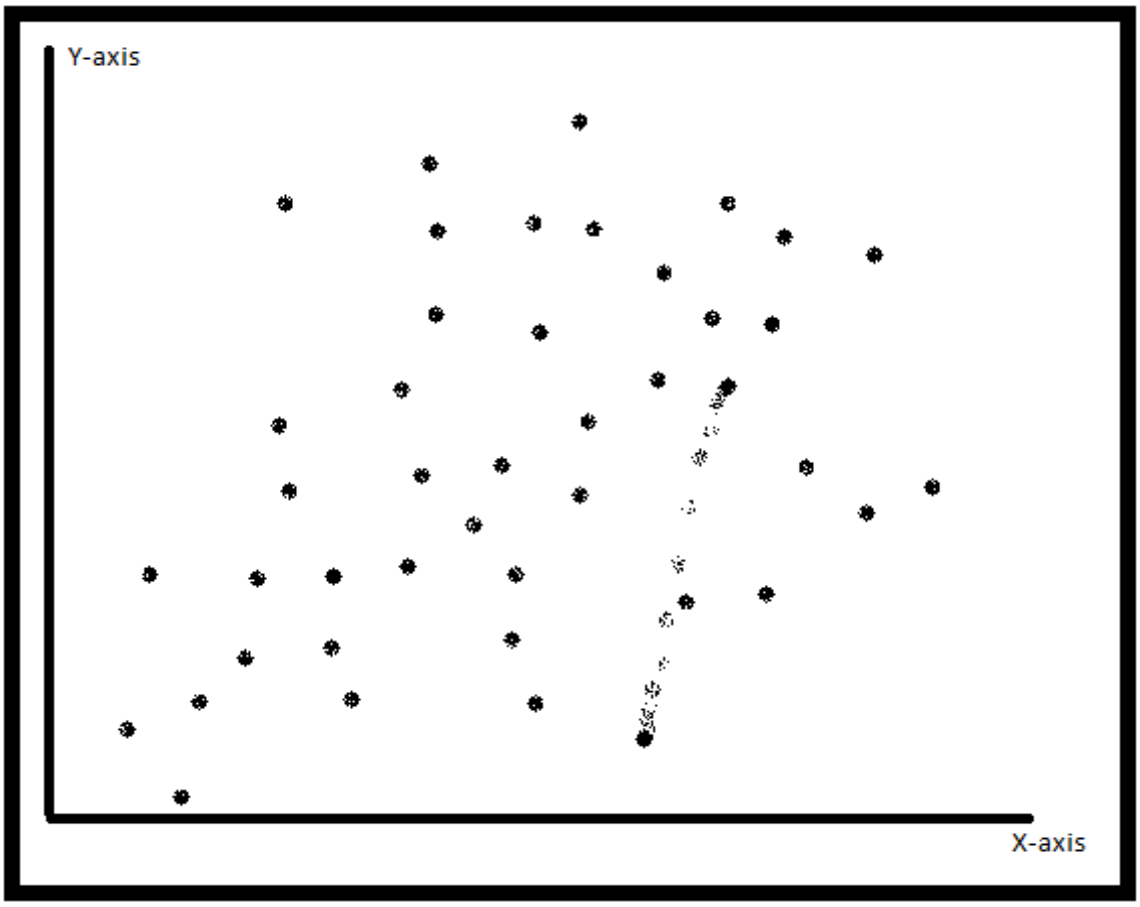


**Fig. 1.7 Random data on 2-D plane**

Figure 1.7 illustrate that data is arranged randomly on 2-D plane. Next step is to do clustering. For clustering basic clustering algorithm will be followed. K-mean is the basic clustering algorithm. K-mean clustering algorithm is based on partitioning clustering technique. As explained below.

**1.5.2 K-mean Clustering: -** K-mean clustering is based on partitioning based clustering. For performing K-mean clustering D data set is selected which contain n number of objects in D data set. First of all data is divided into k partitions and assign a center element to each

partition. K-mean clustering usually performed with low dimensional data sets. K clusters assign n data objects that are similar to each other within a cluster but dissimilar to each other outside the cluster. Algorithm firstly select k clusters randomly from D data set which describes the center of the clusters or say mean of the clusters. After assigning center to each cluster on the basis of similarity add more data items into the cluster on the basis of Euclidian distance between the objects. After adding more objects to a cluster find a new mean or say center of cluster by calculating the mean of the objects that are present in any cluster and then add more objects to a cluster on the basis of new center. This process iterates until no changes occur to the assignment of objects.

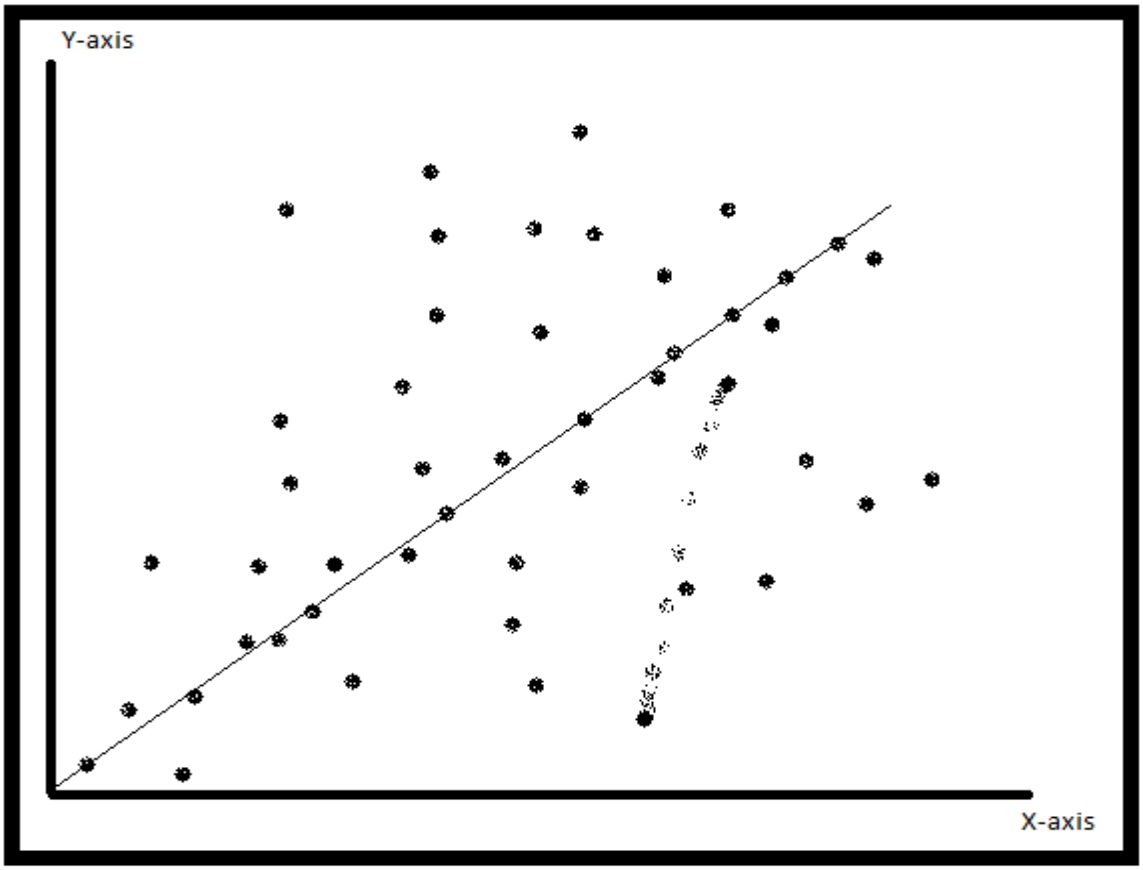


**Fig. 1.8 K-mean clustering on 2-D plan**

Figure1.8 illustrates that arranging data in 2D plane. Categorized data is used to apply k-mean clustering on the basis of the similar properties. Data which have similar properties arranged in one cluster and other is arranged in another cluster.

**1.5.3 Spectral Clustering:** Spectral clustering is a method of density-based clustering in which all the data points of the data set are characterized in the form of nodes and the interactions among them are denoted with the help of edges that convey some weight. High weight shows high similarity among the two nodes and the low weight is a suggestion of the distinction between the two nodes. Weights stuck between the nodes are documented in the form of a matrix and the matrix so created is known as distance or weight matrix. This distance matrix is then transformed into resemblance matrix using some alteration method. After this, laplaceian matrix is constructed which is usually the difference between the distance matrix and the similarity matrix, the succeeding step is to calculate the Eigen values and construct a matrix that contains these Eigen values in the form of columns. Then cluster the data points using some clustering algorithm.

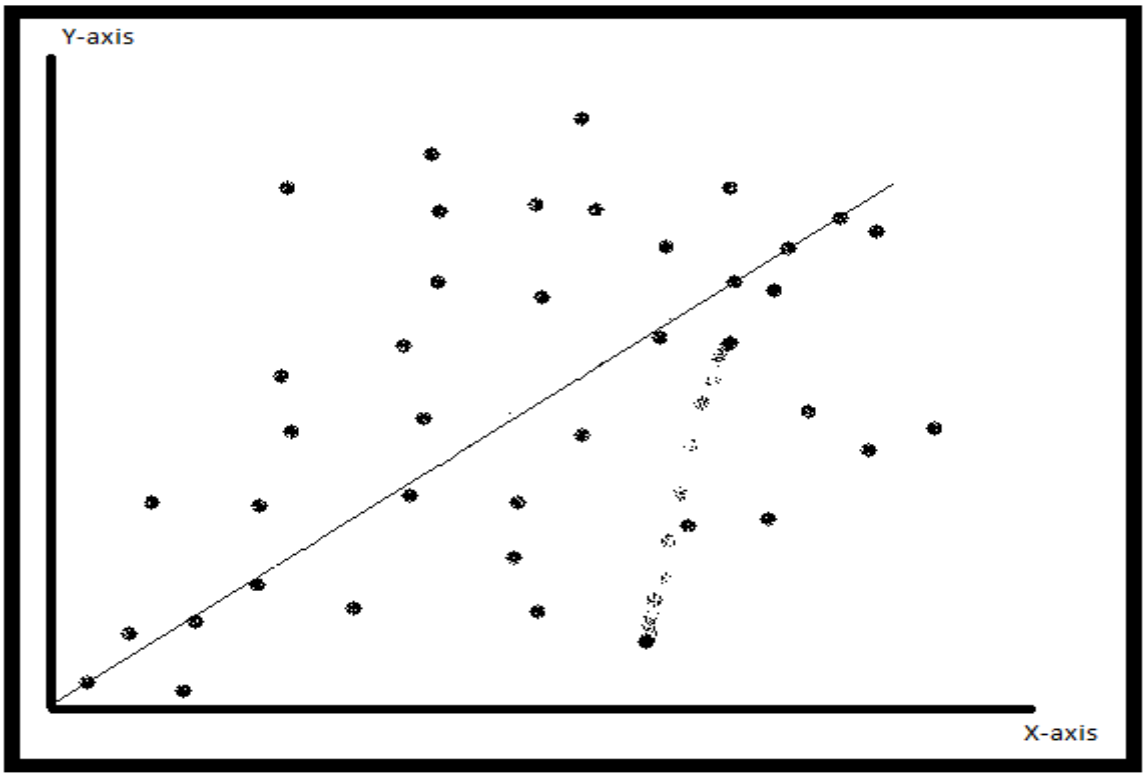


**Fig 1.9 Spectral Clustering**

Fig1.9 illustrates the arrangement of data set in 2D plane and then applies k-mean clustering on the data set. K-mean is basic clustering using this some limitations are there to remove these limitations spectral clustering is applied on data set.

**1.5.4 Adjustment Factor: -** Adjustment of the points is done to improve the cluster quality. Cluster quality is improved by adjusting the number of points present in two D plane.

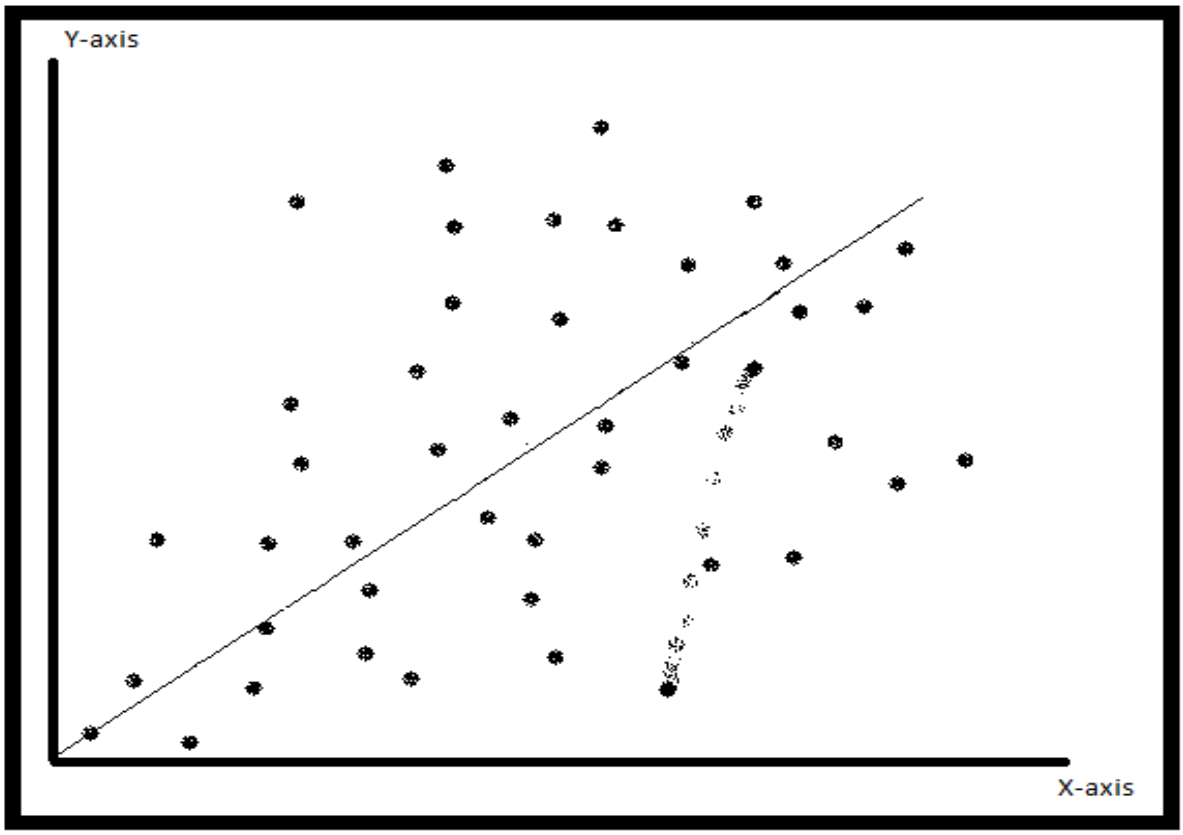


**Fig1.10 Adjustment factor**

Fig 1.10 illustrates the adjustment factor. Here adjustment of points is done just for the sake of improving cluster quality. Cluster quality is improved by adjusting the number of points in given two dimensional data as shown in figure.

**1.6 Mixed data clustering: -** Approaches that are used to join numerical and nominal data variables and distance are inspected among the variables. Before joining the overall distance measure separate distance measures are calculated for numerical and nominal data variables. All insignificant variables are converted into or improved into numeric variables. After that separation measurement is done using all present variables. Ability of the algorithm is measured that how it can handle the all data variables. Consistency and correctness of the algorithm is measured and equated with other algorithms or say data mining algorithms.

Multivariate clustering algorithm always uses data set's variables to find the similarity between the data items. Field variables are used to build clusters. Field variables are variables i.e. all variables except spatial variables and time variables. All field variables contain pressure, density and temperature.

**1.7 Applications of Clustering: -** There are many applications of Clustering in data mining and other fields some of them are listed below. Basically two applications of cluster analysis are as given below.

**1. Summarization: -** Clustering reduces the actual size of data set and data becomes in aggregated form and more concise and understandable contains less errors more accurate i.e. high quality data.

**2. Understanding: -** Browsing, Group gene and proteins are grouped together which contain similar functionality.

**Application of clustering in data mining**: - Clustering is one step of data mining process. It groups the similar objects into a single cluster. Population segmentation model is developed by this clustering technique. For example a company that sales many products use clustering to know about the sales of each and every product. Using clustering company become knows about i.e. which product's sale is growing and which one is lacking.

**Clustering is also used in text mining**. It is discussed as a text data mining in text analysis. It include the structuring the input text and derive the pattern through these structures finally evolution and interpretation of output data. Typical text mining task contains usually clustering, categorization, extraction, sentiment analysis, entity relation modeling. Extraction information from hidden patterns in big text data collection called text mining. These were the basic clustering applications that are discussed here some of the other clustering applications are there which are discussed in next section.

1. **Weather report analysis: -** Clustering is used in many applications means real time applications many real life examples of clustering are there to cluster the data.

Weather report analysis is one of the many applications of clustering. Using clustering one can easily know about the weather report of the future day.

2. **Image processing: -** Image processing is another field in which clustering is used to cluster and process the images. It is very useful in image processing field. It provides the homogeneity in the images.
3. **Machine Learning: -** Clustering is very useful in machine learning. Machine learning is basically originates from computer science.
4. **Voice mining: -** For voice mining clustering is used. It is used in various applications like speech recognition and many more fields that use the clustering in them to cluster the voice.
5. **Text mining: -** Clustering is also used in text mining. . Extraction information from hidden patterns in big text data collection called text mining. These were the basic clustering applications that are discussed here some of the other clustering applications are there which are discussed on next page.
6. **Data mining: -** clustering is used to cluster the data sets. Similar data are combined together into a single group. Population segmentation model is developed by this clustering technique. A company that sales many products use clustering to know about the sales of each and every product.
7. **Pattern recognition: -** Clustering is used to find the similar patterns in the data sets.it helps in finding similar patterns in data and match the patterns then group them into a single cluster.
8. **Image analysis: -** Clustering is also used in image analysis and image processing. Clustering analyses the images and find the homogenous images collect them and cluster into single group similarity wise.
9. **Bioinformatics: -** This field is also use clustering to find the information about the botanic terms.
10. **Web Cluster analysis: -** For web cluster analysis clustering is used to analyze the web data and clusters produced from web data. Web data is very large in volume it is very difficult to find the similarity in web data and find interesting patterns in web data is also difficult by using

# CHAPTER-2
# LITERATURE SURVEY

**Literature survey**

This review has been carried out to study different types of clustering techniques to cluster different types of data sets.

**Ashish Patel et.al [1]** presented KL-Divergence with density based clustering to handle uncertain data. Data is associated with uncertain values due to some reasons e.g. erroneousness, sample old-fashioned data sources. These uncertain data are produced daily. Inexact data are difficult to handle. These inexact data include moving items. These moving objects include vehicle or people. This is based on location. Inexactness is depends upon location of data items. Probability distributions are used to describe these inexact data items. Data items are clustered on basis of probability distributions. It is very challenging task that clustering of inexact data items. In early days partitioning method and density based clustering methods were used to cluster inexact data. These methods face problems in handling uncertain data because items are indistinguishable. By using probability distribution uncertain objects can be modeled uncertain data objects. Purposed paper describes Kullback-Leibler divergence method to describe the similarity between data items and integrate that data items into partitioning and density based technique. In initial stage find uncertain objects then cluster uncertain data items using partitioning technique and remaining data clustered using traditional clustering techniques.

**Ahmed Rafea et.al [2]** aimed at topic extraction in social media by using bisecting K-Mean clustering algorithm. Any word that occurred more than 20 times will used in whole corpus for clustering the tweets and hot topics into a single group or cluster by using bisecting K-mean clustering.

**Alvaro Garcia-Piquer et.al [3]** presented an evaluation of cluster representation of multi-objective clustering. These are based on many optimization objective functions which are

evolutionary in nature. There capabilities allow them to select a better solution than conventional clustering algorithms.

**Alessia Albanese et.al [4]** gave a technique to extract knowledge from spatial-temporal data. Rough set technique to spatiotemporal data for outlier detection is used. ROSE approach is used to detect outlier and extract useful information from spatiotemporal data.

**Aliya Edathadathil et.al [5]** presented modified K-medoid method to cluster uncertain data objects on the basis of probability distributions similarity. Previous studies used k-means, UK-mean and density based clustering algorithm to cluster data objects that are geometrically distinguishable but they can't handle objects that are indistinguishable. Now KL-divergence with K-medoid used to cluster uncertain objects that are geometrically indistinguishable.

**Anirban Mukhopadhyay et.al [6]** gave a survey on multi-objective evolutionary algorithm for data mining. Firstly multi-objective evolutionary algorithm used for feature selection and classification. He presented different multi-objective evolutionary algorithms for clustering, association rule mining and many other data mining tasks.

**Ana L.N. Fred et.al [7]** aimed at evidence accumulation for combining multiple clustering. Data set partitioned into multiple partitions. These partitions can be made using different clustering algorithms and using same clustering algorithms. These partitioned can be combined using Evidence accumulation concept based on voting mechanism. Among n patterns a new similarity matrix generated i.e. n into n matrix. Hierarchal clustering is used to obtain final data partition of n patterns. Split and merge approach is applied for combination of multiple clustering.

**ArnostKomarek et.al [8]** presented a clustering technique which cluster data i.e. multivariate continuous and discrete longitudinal data. Method of classification is described as on the basis of longitudinal measurements. Basically clustering procedure depends upon the classical GLLM given for each marker. Bayesian approach used for the classification or say clustering purpose here. Continuous and discrete data clustered easily using Bayesian classification approach.

**ARISTIDES GIONIS et.al [9]** presented clustering aggregation technique to aggregate the clustering technique to cluster data efficiently. Clustering aggregation is done for sake of solving some clustering problems like clustering categorical data. Clustering aggregation can be viewed as meta-clustering. This is clustering of clustering is called as meta-clustering. Large data sets use sampling to scale the algorithm

**Bin Jiang et.al [10]** presented clustering uncertain data based on probability distribution similarity. Previous method used extended partitioning clustering and density based clustering to cluster uncertain data. Data can't be efficiently clustered using these approaches because of uncertainty. Probability distribution uses KL-Divergence method to find the similarity between uncertain objects and after that apply density based clustering technique to cluster remaining data. Effectiveness, efficiency, scalability of this extended approach is greater than previous clustering techniques.

**Bo Liu et.al [11]** presented an efficient framework of one class learning and concept summarization learning on uncertain data. Solution is divided into two subparts firstly one class learning to handle uncertainty and second support vector based clustering to cluster or summarize the concept.

**Cane Wing-Ki Leung et.al [12]** aimed at probabilistic rating inference framework for mining user preferences from reviews after mining user preferences map those into a numerical rating scale. Linguistic processing techniques are used to extract words from reviews. Sentimental orientation and strength of opinions are determined using relative frequency based method. Collaborative filtering is used to filter the data that is important.

**C.Deepika et.al [13]** presented An Efficient Uncertain Data Point Clustering based on Probability-Maximization Algorithm in which partitioning method and density based clustering is used to cluster the uncertain data. Partitioning method contains k-mean clustering and density based clustering contain DBSCAN that are used in proposed method to handle uncertain data or say clustering of uncertain data. In this purposed algorithm data sets are firstly preprocessed and in preprocessing step basic components are sizes, classes, attributes, standard deviations. After preprocessing step probability maximization algorithm

is performed. In PM algorithm two partitions are there one is true partition and second is clustering results

**D. M. PAdulkar et.al [14]** presented VORNOI diagram and R-tree with Ensemble for clustering uncertain data. Clustering uncertain data is very difficult task. Uncertainty of data is just because of precision of device that is used for checking readings. Mostly data is uncertain which is collected from satellite, sensor network and many more resources may contain uncertainty. Clustering these types of data sets using VORONOI diagram and R-tree is not difficult. Data can be easily handled using VORONOI diagram. It would produce efficient results of clusterings.

**Dimitrios Mavroeidis et.al [15]** presented K-mean clustering stability it is because of feature selection. Clustering stability is taken into account and discovers the effect of maximization of stability in continuous K-mean clustering problems. Mathematical optimization and statistical analysis performed here that describe the algorithm's stability. The purposed algorithm is basically based on sparse PCA approach which increases the stability using greedy approach on the basis of feature selection. Practical implementation of proposed work is in cancer research

**Francesco Gullo et.al [16]** gave uncertain data clustering using uncertain centroid based partitional clustering. Cluster centroid base partitional clustering play an important role in cluster uncertain data. In terms of random variable cluster centroid is seen as an uncertain object. This gives better presentation of clusters. Clustering performance improved and maintains equal productivity with already present system.

**Fatih Dikbas et.al [17]** presented K-mean clustering method for defining homogeneous regions for streamflow processes. This technique is used for flood estimations. K-mean algorithm is used for estimation maximum annual flow in a particular area and also described about hydro logically homogeneous groups. Seven regions are identified by clustering analysis. K-mean method is used for flood and frequency analysis.

**Feng Cao et.al [18]** presented a density based clustering over an evolving data stream with noise. Clustering plays an important role in evolving data stream. Most of the time demand for clustering is number of clusters, handling outliers and discovering clusters with arbitrary

shape. A new approach of discovering clusters in an evolving data stream i.e. Den-Stream it produced clusters of arbitrary shape and handle outliers easily.

**F.U. SIDDIQUI et.al [19]** aimed at optimized K-means clustering algorithm for image segmentation. Optimized k-mean clustering algorithm segments an image homogeneously into region of interest with ability of escaping dead center and trapped center. It produces more homogeneous segmented images and is effective in avoiding dead center and trapped center at local minima.

**G. W. Ma et.al [20] presented** an enriched K-mean clustering algorithm for cluster fractures with meliorated initial centers. Hierarchal clustering is used to select the initial cluster center. Synthetic data is used to describe effectiveness of given algorithm. Density based method is also useful in gathering degree concept to find the initial cluster center. Algorithm can be easily implemented for geographical mapping and its convergence is also very fast.

**Graham Cormode et.al [21]** gave approximation algorithm for clustering uncertain data, uncertainty arise in applications like sensor networks measurements output of mining algorithms and methods. There are many bacteria approximation algorithms. For best k-center one can achieve the approximation by picking center. Another achieves a constant factor approximation that pick 2k center. For clustering uncertain data these results are combined.

**Jean Christoph Jung et.al [22]** presented a promising method for managing data that is extracted from web. Extension of OBDA is presented over here that capture uncertain data with the help of probabilistic data model. It replaces the uncertain answers with the certain answers with the help of probabilistic computations that are more certain than that before. New approach relates to probabilistic database system i.e. PDBMs in same that OBDA relate to RDBMS. At last one can say that by using ontology uncertainty of data can be handled and data is managed by ontology. Means to say that ontology has many applications in uncertain data management

**LipikaDeyet.al [23]** presented opinion mining from noisy data. Data on web is very huge data and to extract important data from opinions. Data on web may contain errors i.e. grammatical mistakes, spelling mistakes, punctuation problems and irrational capitalization.

Opinions are aggregated at many levels i.e. product level or say feature level, site level. A system is developed based on this approach which analyzes opinions taken from web blogs.

**Markus M. Breunig et.al [24]** gave a local outlier factor which identifies the local outliers presented in data sets. These outliers are density based local outliers. This approach is used for detecting criminal activities and finding rare instances. Each object in data set is assigned a degree of being an outlier this degree is called as local outlier factor i.e. LOF. Outliers are removed from data and remaining data remains as clusters data

**Michael Chau et.al [25]** presented UK-Mean clustering algorithm to cluster location data i.e. uncertain data which is not certain like moving objects. Data contain inherent property i.e. uncertainty it considered to be required high quality results. UK-Mean clustering algorithm produces good results when applied on uncertain data.

**Patrick Glenisson et.al [26]** presented meta-clustering of gene expression database and some information that is based on literature. Many biological gene expressions are clustered using meta-clustering. In this combined analysis of literature base information and expression data that give meaningful clusters of data but not give efficient results when microarray used alone. The joint analysis is meaningful

**Reshma MR et.al [27]** gave a good technique that can help to manage uncertainty and cluster uncertain data based on KL-Divergence technique. Many problems in data are comes due to uncertainty of data. From those problems clustering is one of them. Due to this uncertainty of data there are problems in clustering. Previously known methods i.e. portioning method and density based method adopted to handle uncertain data and cluster that uncertain data into a single cluster that is the data items in one cluster are similar to each other. Portioning method plus density based methods are that which use or say based on geometric distance between objects. In purposed algorithm probability distributions are used which are the mandatory features of uncertain objects. In purposed algorithm Kullback-Liebler divergence is used. This algorithm is basically for measuring the resemblance among objects and integrates portioning plus thickness based method to group inexact objects. FCM method is used in this to cluster uncertain data objects and show effectiveness of the data objects. FCM means fuzzy c-mean clustering for data with tolerance. Data objects are

represented by probability distributions. And probability distributions is described by probability mass function and objects that's are in continuous distributions are described by probability density function. In this paper basically KL-divergence is integrated with portioning and density based clustering to handle uncertain data and clustering of uncertain data

**Rich Caruana et.al [28]** gave meta-clustering to cluster many clustering techniques. There are many clustering techniques that are used to cluster data. Everyone search for best clustering technique which fit their needs. Meta clustering finds variety of clusterings and then clusters these clusterings. Clustering using Meta clustering produces efficient results.

**Samir N. Ajani et.al [29]** presented improved K-mean clustering algorithm for clustering uncertain data objects. Improved K-mean clustering algorithm use indexing which help in clustering uncertain data. It minimizes the computation time and handles uncertain data objects. Usually clustering on uncertain data is very hard. It is very difficult to handle. In previous studies density based and partitioning clustering algorithms are used to handle uncertain data. They can't handle uncertain data efficiently hence improvement in k-mean is applied.

**Tina Eliassi-Rad Terence Critchlow et.al [30]** illustrated a cosine similarity technique which helps the scientist to discover new knowledge from simulation. Cosine similarity technique reduces the modeling time. Mesh format used in this technique which perform the multivariate clustering. Mesh data is there which varies with the time. Mesh data contain multiple dimensions in it. List of jones are used as input and list of clustered are produced as output of data items. Red and green colored data items are available here red colored data that is already present in cluster. Green data item ready for clustering. Linking algorithm is used to describe the proper location of the cluster. Different clusters are linked together with the help of linking algorithm. All red data items are clustered into clusters or grouped into clusters. Clusters are shown as nodes in given technique.

**Tayfun DOGDASet.al [31]** presented a GIS visualizing and EM clustering method for document clustering by using simple multidimensional projection method for visualization and data reduction. Expectation maximization (EM) and partitioning clustering technique i.e.

K-Mean used for transformed data. System customs the Microsoft spatial database engine as a GIS (geographical information system) tools for visualization. Simple multi projection method used cluster the similar data into a single group.

**T HITENDRA SARMA et.al [32]** presented single pass kernel K-mean clustering algorithm which is used to cluster data sets very quickly. In unsupervised classification kernel k-mean clustering perform better than that of conventional clustering algorithms. Space time complexity of these methods is very high, which is not good. Because of this kernel k-mean method is not applied to work with large data sets. Single pass kernel K-mean clustering algorithm is used for large data sets. Its space time complexity is not high. It can be easily implemented for large and complex data sets.

**Weifang SHI et.al [33]** presented environmental risk zoning of chemical industrial areas using k-mean clustering. Environmental risk zoning methods are used to cluster environmental risks. Index quantification model is used to zoning environmental risk by analyzing mechanism of risk happening. By calculating air risk field index, source risk index, water risk field index and then use k-mean clustering to find the environmental risk in the area. Then mapping of zoning results takes place using geographic information system. Zoning is important in risk management and is very useful for decision maker to divide limited resources into sub-areas.

**Xiaoli Cui et.al [34]** aimed at map reduce for optimized big data K-means clustering. K-mean clustering algorithm is very simple algorithm for cluster data. Now a day's data volume is increased, it is quite difficult to cluster big data using simple k-mean clustering algorithm. New model is purposed i.e. map reduce which is novel processing model it eliminates the iteration dependencies and improve the performance. It is more efficient robust and scalable.

**Yi Ma et.al [35]** Presented K-mean and expectation maximization algorithm for model estimation and data segmentation. Probabilistic distributions only use these models and algorithms. First of all model estimation is performed and then data is divided into small parts i.e. data segmentation. Compression and coding is used to segment the data. Segmentation decoupling is also performed here in this technique. Bottom-up approach is

used to obtain the optimal segmentation solution instead of using top-down approach. Lossy data compression uses greedy algorithm.

**Yi Zhang et.al [36]** presented multiple Consensus Clustering to cluster data accurately. He presented consensus clustering and extended Meta clustering to the clustering problems. Consensus clustering finds the one optimal clustering from the given set of clusterings for producing an optimal solution of clustering to cluster data. Meta clustering combines the different clustering into one single clustering and produce one optimal solution to cluster data. Meta clustering and consensus clustering are combined together to produce accurate and efficient results of clustering to cluster the data set. Meta clustering is clustering about clustering. It contain three steps first step generate basic clustering. Combine the all clusterings and produce one optimal clustering to cluster data.

**Zhaohong et.al [37]** aimed toward the soft subspace clustering. Subspace clustering clustering technology to identify clusters based on their associations with subspace in high dimensional space clustering can be classified into 2 different groups' i.e. hard subspace clustering (HSC) and soft sub space clustering (SSC). HSC is been extensively studied by scientific community. SSC are relatively new but more attention on them due to better adaptability. Comprehensive survey on existing SSC algorithm and recent developments are presented over here. SSC has mainly three types conventional subspace clustering, independent SSC, Extended SSC.

**Z. Volkovich et.al [38]** gave a self-learning K-mean clustering for global optimization. Distance is basic factor in many real world learning tasks. Commonly used methods are not appropriate in comparison to distance metric method. K-mean clustering is used to rescaling data and also clustering is done using k-mean clustering so that data separability grows iteratively. Global optimization problem is solved in this. Cluster validation characteristics are used to find the weight matrix.

# CHAPTER-3
# PRESENT WORK

## 3.1 PROBLEM FORMULATION

Clustering and cluster analysis is one of the traditional and more important topic in the field of mining patterns. It is the initial stage in the direction of stimulating in formation detection. Clustering is development or say procedure of assemblage data items into groups of separate classes, that groups are named as clusters. Now items inside a class or say within a cluster are having high resemblance to each other in the meantime objects in separate classes or say in different clusters are more unlike. Clustering is a method or say process to assemble related papers, but it fluctuates from classification of documents are gathered on the coasting place of through the use of already defined topics. Clustering has numerous applications. One more additional benefit of clustering is that documents appear in numerous subtopics, thus guaranteeing that a beneficial document that is require would not be misplaced from examine outcomes. A basic clustering algorithm produced a vector of topics for each document and measures the weights of how easily and how healthy the document fits into each cluster. In this today's world, whole data is digital data. Due to digitalization much of the data produced is uncertain which cannot produce the relevant data for the organization. In day to day life many data is produced daily. More amounts of data produced are digital data, and such digital data is uncertain in nature. Handling such uncertain data is very difficult task. Distance between uncertain object is described by one numerical data value. Clustering of uncertain data is very difficult task and is also challenging task too. In early days extended portioning method k-mean plus density based technique DBSCAN were used. Inexact objects are model in discrete and continuous domain. In purposed paper Kullback-Leibler divergence technique is used to quantify resemblance among in exact objects in both region i.e. discrete and continuous domain, and then integrate dividing plus density based clustering algorithms to cluster inexact data objects. Firstly uncertain data objects have been found and then cluster uncertain data using partitioning method and then remaining data is clustered using traditional method of clustering. In the proposed algorithm the problem of accuracy can be raised. As the centroid of the cluster will not be exactly defined this increases the cluster

quality. In this work, enhancement is proposed to increase the cluster quality and increase the accuracy and efficiency of the algorithm

## 3.2 OBJECTIVES

1**.** To study and analyze density based clustering techniques for cluster uncertain type of data

2. To propose enhancement in the density based probability modal to increase cluster quality for uncertain data clustering in Data mining

3. The enhancement will be based on the modification of Probabilistic function for density-based clustering

4. To implement proposed technique and existing technique and analyze the performance in terms of cluster quality and accuracy

## 3.3 RESEARCH METHODOLOGY

In data mining various clustering algorithms are used to cluster data. Maximum problem occurs in probability distribution clustering algorithm is of accuracy. The cluster quality is not exact good by using conventional clustering techniques i.e. density based clustering and partitioning clustering. To improve cluster quality and to handle uncertain data i.e. digital data KL-divergence algorithm is used. Now for further enhancement in quality and accuracy further enhancement in the clustering algorithm for improving cluster quality is required. In given methodology first of all data is loaded into a single file. After loading the data set sigma value is initialized and then apply KL-divergence method for finding similarity between data items. Then apply density based clustering on remaining data. Uncertain data are clustered using KL-divergence method, 80% of data can be handled using this. Remaining 20% data is grouped into a single group using density based clustering. Normalization is applied after applying these two techniques for clustering. After applying normalization data is clustered into clusters. Flow chart of given methodology is given below. Proposed methodology flow chart is also given below in fig 3.2, in this proposed methodology data is loaded in iterations Meta clustering and K-mean clustering is applied to cluster the uncertain data items. Normalization is applied again and again to cluster data efficiently. Data is clustered using Euclidian distance. It is used to find similarity.

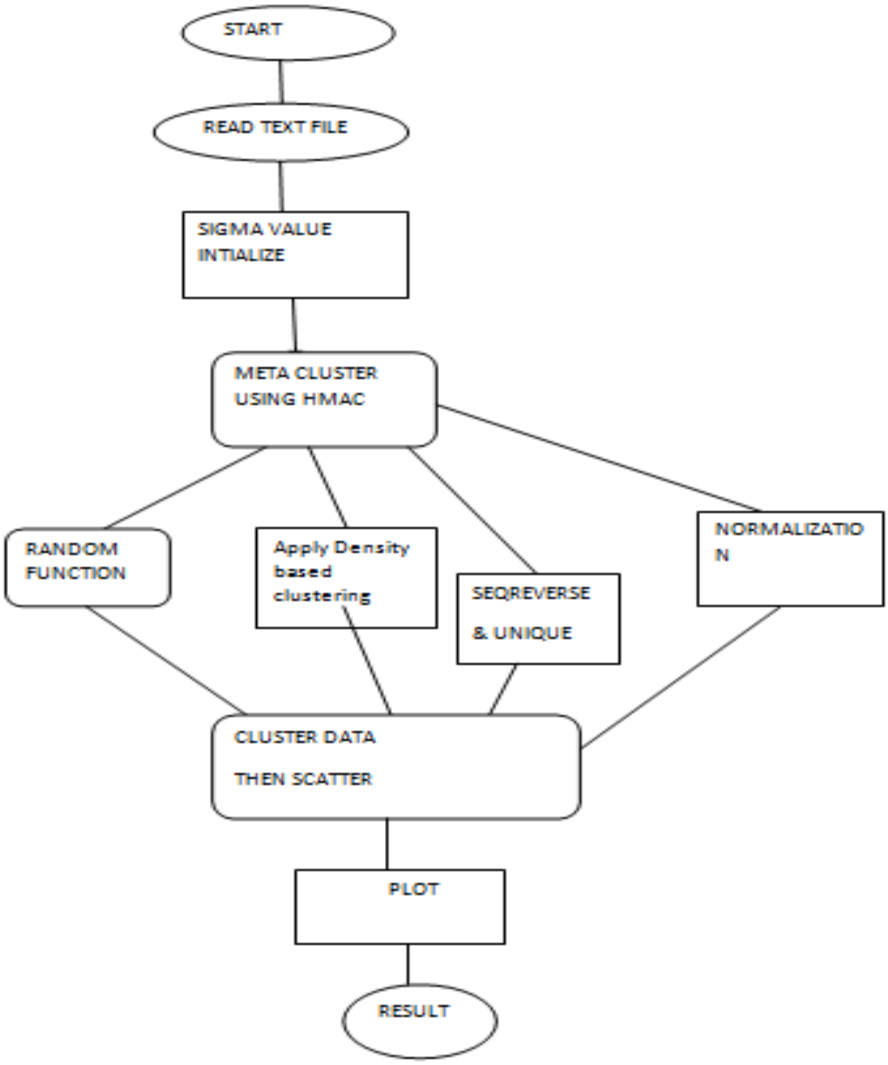


**Fig 3.1: Basepaper Flowchart**

As illustrated in the figure 3.1, the procedure of meta-clustering is shown. In this flowchart functions which are implemented works in following manner.

1. **Load dataset and Sigma value Initialization:** In the first step of the flowchart, the dataset will be initialized on which the meta-clustering has to be performed. After loading the dataset, the sigma function will be applied on the dataset. In the sigma function, the initial points are selected from the dataset on the bases of probability to select centroid for meta clustering
2. **Meta clustering using HMAC:** In the second step, HMAC function is applied on the loaded dataset. The HMAC functions stands for hierarchal Meta clustering. In this functions whole dataset and dataset is divided into hierarchal order. In which the

similar elements form one cluster and so on. The similar elements are marked with the same colors to form hierarchy

3. **Random function and Density Based clustering: -** In this step random function is applied to select point randomly from the dataset for the normalization. When normalization point is selected from the dataset. The density based clustering is applied from the dataset. The density is calculated from each hierarchical point. To cluster the data from random dataset, density based clustering is applied.
4. **Unique and normalization: -** The dataset which is loaded firstly hierarchical clustering is performed, after calculating hierarchy of the dataset unique points are calculated from each hierarchy. The unique points are then normalized from hierarchical points. 3NF polynomial Normalization is applied to find the unique points in the data set to cluster data efficiently.
5. **Clustering:** - The point which are unique and are normalized, then the clustering is performed on 2D plane. Each cluster is marked with different colors. After finding the unique points similar data items are grouped together. Clustering is done on the basis of Euclidian distance. It is used to find the similarity between the data points present in two dimensional plane. Formula for finding Euclidian distance is as given below.

   E (Dist) = $((X_2-X_1)^2+ (Y_2-Y_1)^2)$

   This is the basic formula for finding similarity between the objects. Objects that are more similar to each other are closer to each other. Means if distance between the objects is lesser then objects are similar to each other. Euclidian distance is basic building block of clustering in K-mean algorithm. After clustering using Euclidian distance and normalization plot function is used to plot the data clusters into two dimensional plane.
6. **Plot function: -** This is used to draw the clusters into two dimensional plane different colored clusters of data are plotted into two dimensional plane. First of all data sets are scattered randomly in two dimensional plane after using the clustering process data is clustered using given methodology. This was all about the given methodology for clustering uncertain data. Now proposed methodology for enhancement in clustering to improve the accuracy is as given below in fig. 3.2

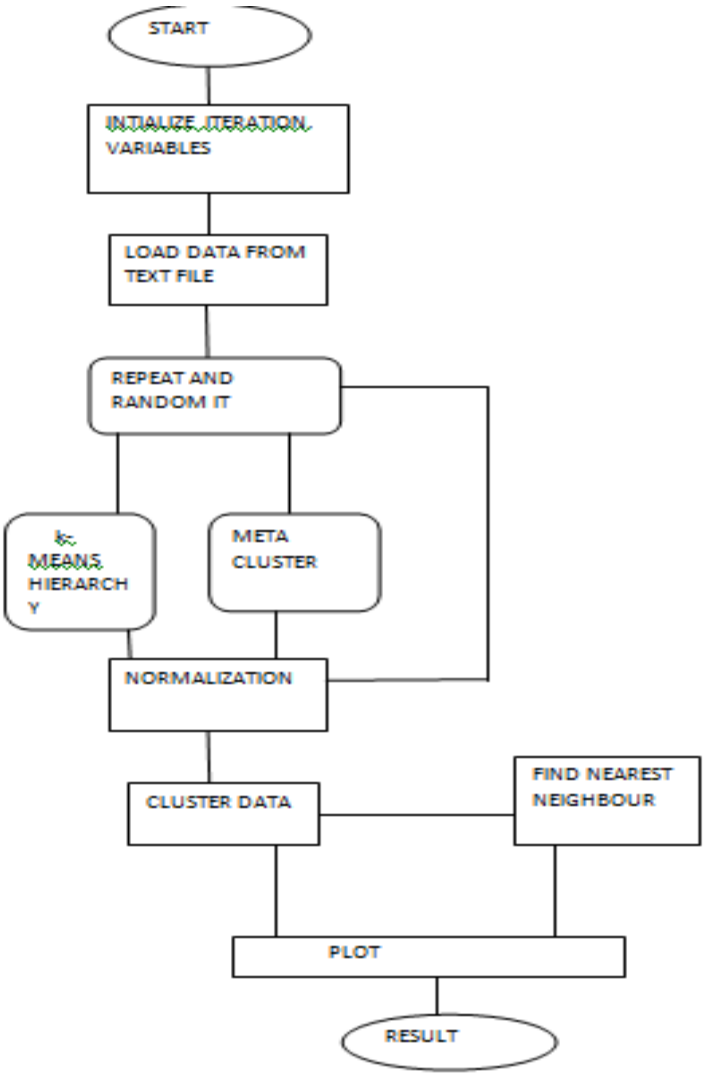


**Fig 3.2: Research Methodology Flowchart**

As shown in figure 3.2, the flowchart of new technique is shown. In the proposed technique the normalization is implemented in the iterative order. In proposed work, the dataset is loaded and from the dataset two values defined first values is of initial centroid point. The second point is of integrations for normalization. From the dataset, random points are selected for clustering the data or to find common data. The K-mean hierarchal clustering is applied to divide the dataset into different parts. The second part is of meta-clustering which is based on KL divergence theorem based for clustering the data. The hierarchal divided data and meta-clustered data which will normalize for number of iteration time which is defined

in the start of the flowchart. The data will be clustered in the last step using the k-nearest neighbor algorithm which is based on Euclidian distance. Euclidian distance plays very important role in finding similarity between the data points. Two data points whose distance is lesser in between them they are more similar to each other. In the end of the flowchart the clustered data will be on the 2D plane. Euclidian distance can be measured by using formula.

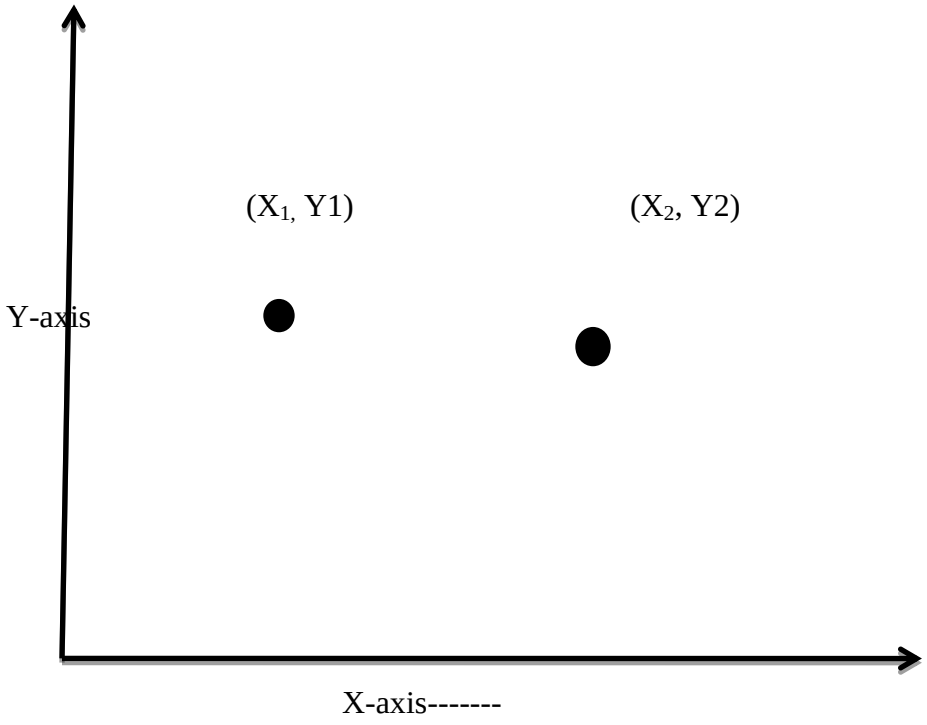


**Fig. 3.3 Two points in 2D plane**

Formula for calculating Euclidian Distance is as given below.

$$E(X, Y) = \text{sqrt}\ ((X_2-X_1)^2+ (Y_2-Y_1)^2)$$

X-axis and Y-axis are given above in this figure. Two points are shown on two dimensional plane. Each point includes X coordinate and Y coordinate in X, Y plane. First point contains X1 and $Y_1$ coordinates. Second point contains $X_2$ and $Y_2$ coordinates. Euclidian distance is calculated using these coordinates as given in formula for calculating Euclidian distance.

# CHAPTER-4
# RESULTS AND DISCUSSION

**TOOL**

The proposed idea will be implemented in MATLAB which is widely used in all areas of research universities, and also in the industry. MATLAB is beneficial for mathematics equations (linear algebra) moreover numerical integration equations are also solved by MATLAB It is also a programming language, and is one of the simplest programming languages for writing mathematical programs.it has various type of tool boxes that are very beneficial for optimization and so on. Mathematical tool is used for implementation of purposed algorithm

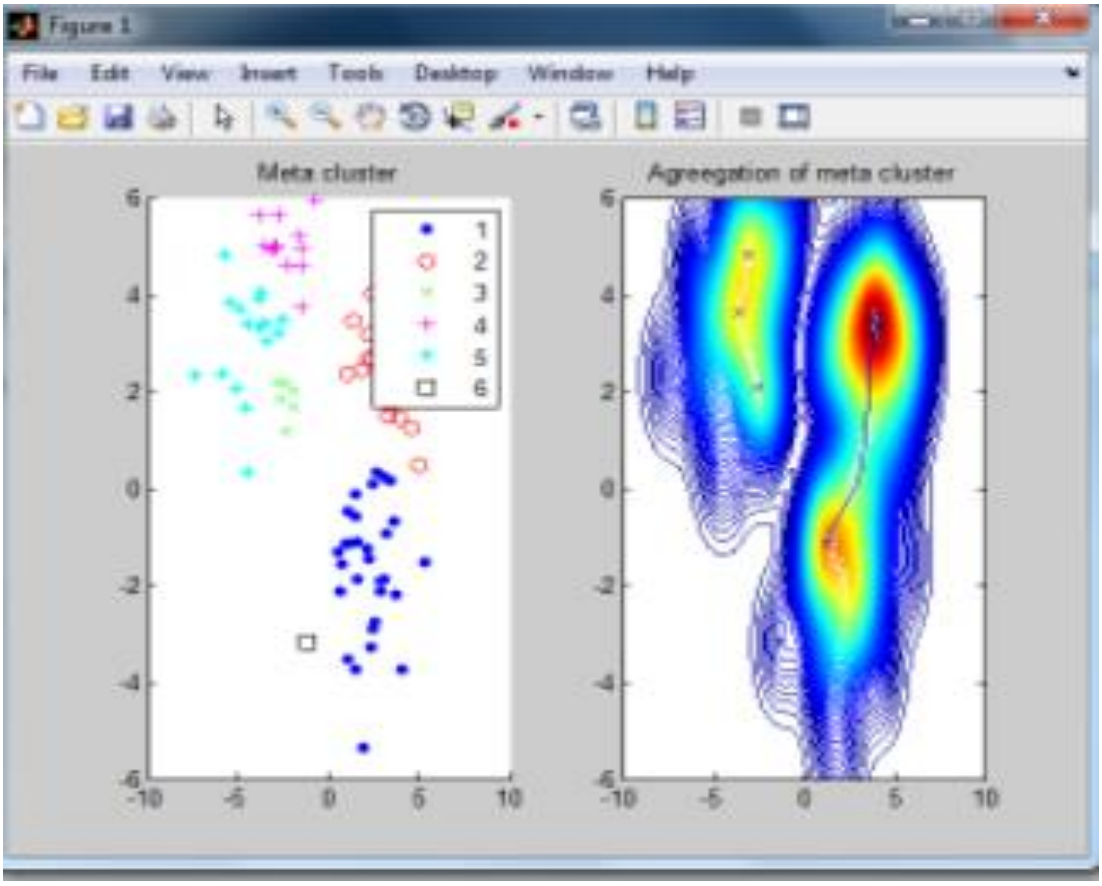


**Fig 4.1 KL-Divergence with Density Based**

As illustrated in figure 4.1, the base code of the Kl-divergence with density based clustering is implemented. In the process of meta-clustering the dataset is loaded which will be normalized. After the normalization HMAC will be implemented to cluster the data

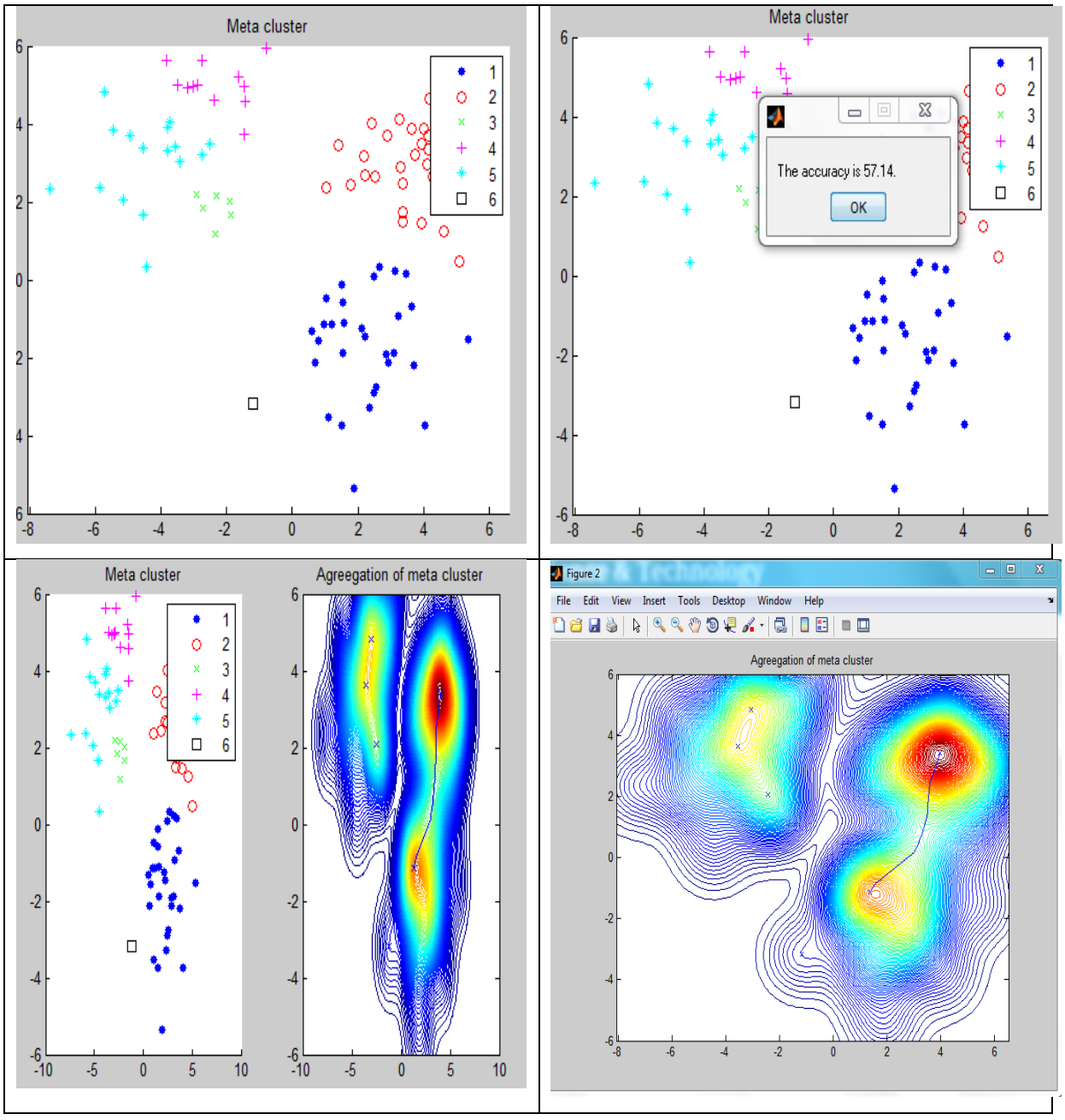


**Fig 4.2: Output of meta-clustering**

As shown in the figure 4.2, the dataset which is loaded first the random point is selected using the sigma function which is shown in the figure with the black square. The data will be aggregated using the density based clustering. The density will be finding out in each

hierarchy using the HMAC. In the last step the clustered data will show with different colors and plotted on 2D plane

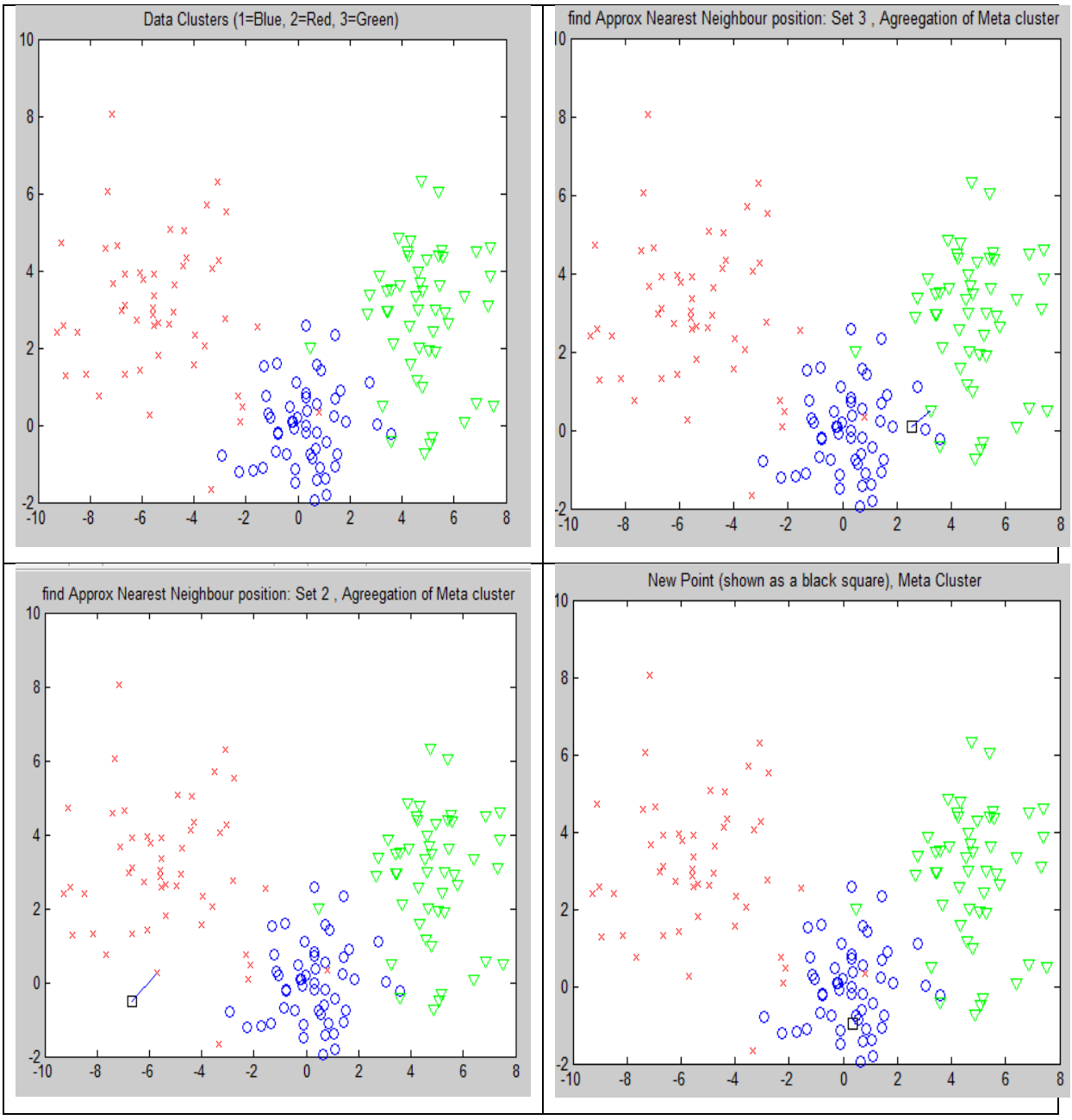


**Fig 4.3: Enhanced Meta clustering implementation**

As shown in the figure 4.3, the dataset is loaded and numbers of iterations are defined for normalization. According to number of iterations the dataset will be clustered on the 2 D plane. After loading the data sets iterations are started. HMAC function is applied on the data set to adjust the data set into a hierarchy. Hierarchal clustering is applied on the data set to

adjust the data sets into a hierarchy. Similar data are adjusted into a single hierarchy. After making the hierarchy of data K-mean clustering is applied on hierarchal data. After applying k-mean base clusters are produced this is based on Euclidian distance of the data elements. Nearest neighbor distance between the elements is calculated and then data is clustered using K-mean clustering algorithm. After clusters are created then apply Meta clustering to further cluster the clusters.

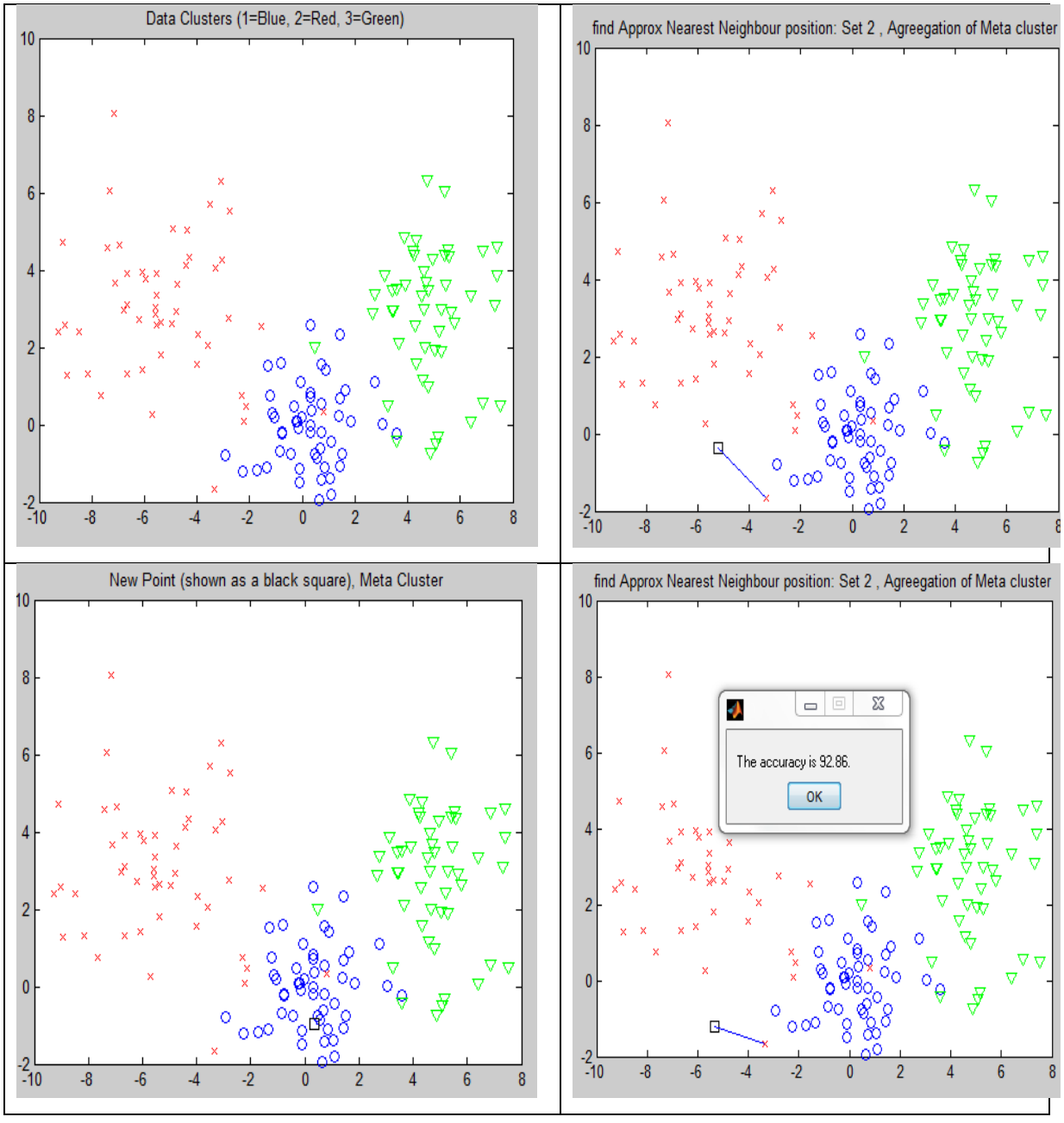


**Fig4. 4: Selection of normalization point randomly**

As illustrated in the figure 4.4, the dataset will be loaded and from the loaded dataset the sigma function is applied. The sigma function will gave the most relevant point value which is used as a central point to cluster the data elements. Sigma function is used to find the most relevant central point on the basis of which data will be clustered easily and more accurately. Accuracy of the clusters is totally dependent on the sigma function. Central point decides whether which data element is added to which cluster and hence accuracy of the clusters is on the basis of central elements. .

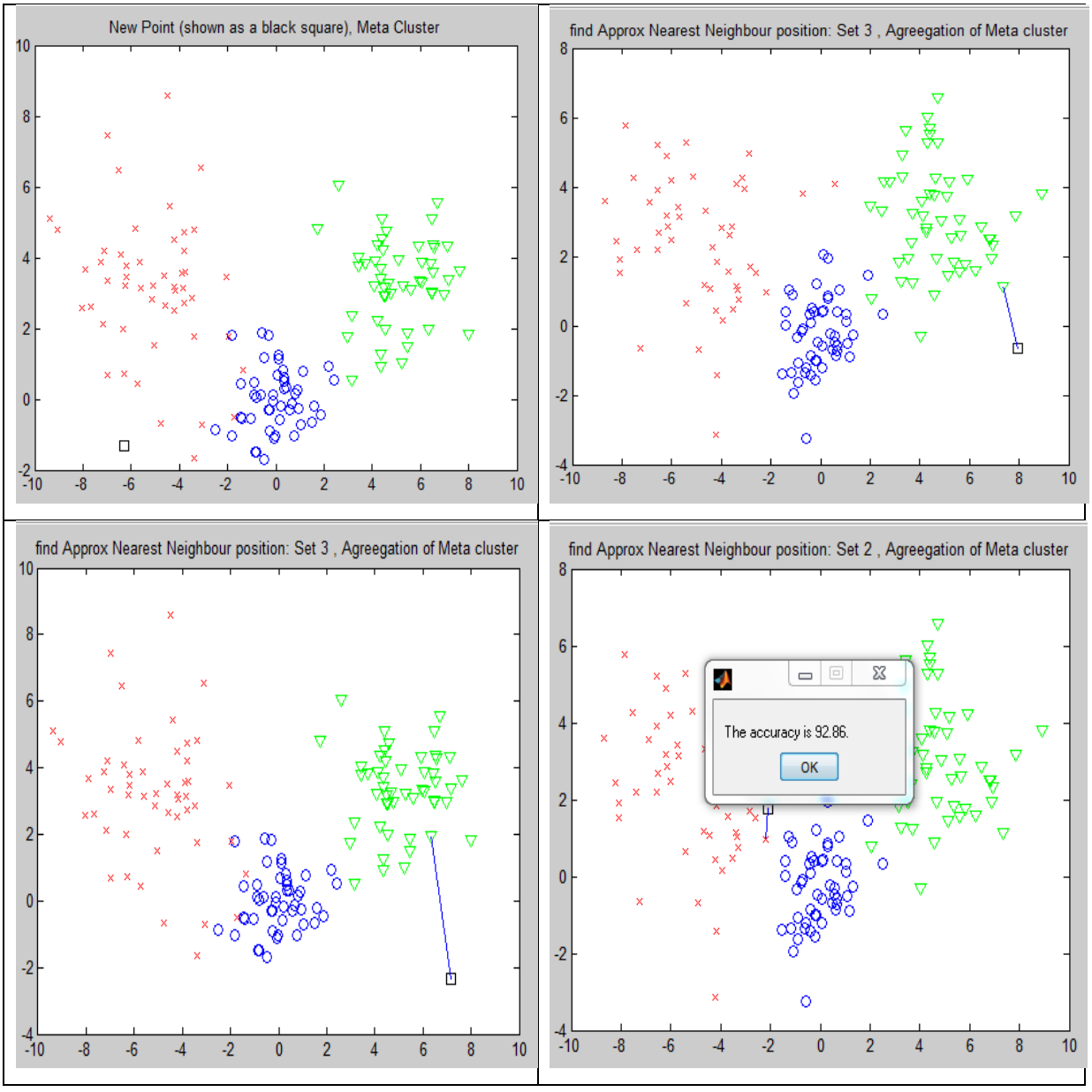


**Fig 4.5: Normalization is applied for central point selection**

As shown in the figure 4.5, the dataset is loaded and on the loaded dataset number of iterations are selected, the HMAC is applied and Meta clustering with K-mean and density based clustering is applied. After clustering the data normalization point is selected for the number of time for clustering the data. Normalization point is used to find the unique elements in data sets, which is used to cluster data elements that are similar to each other. Similar elements are grouped together and dissimilar elements are removed as noise.

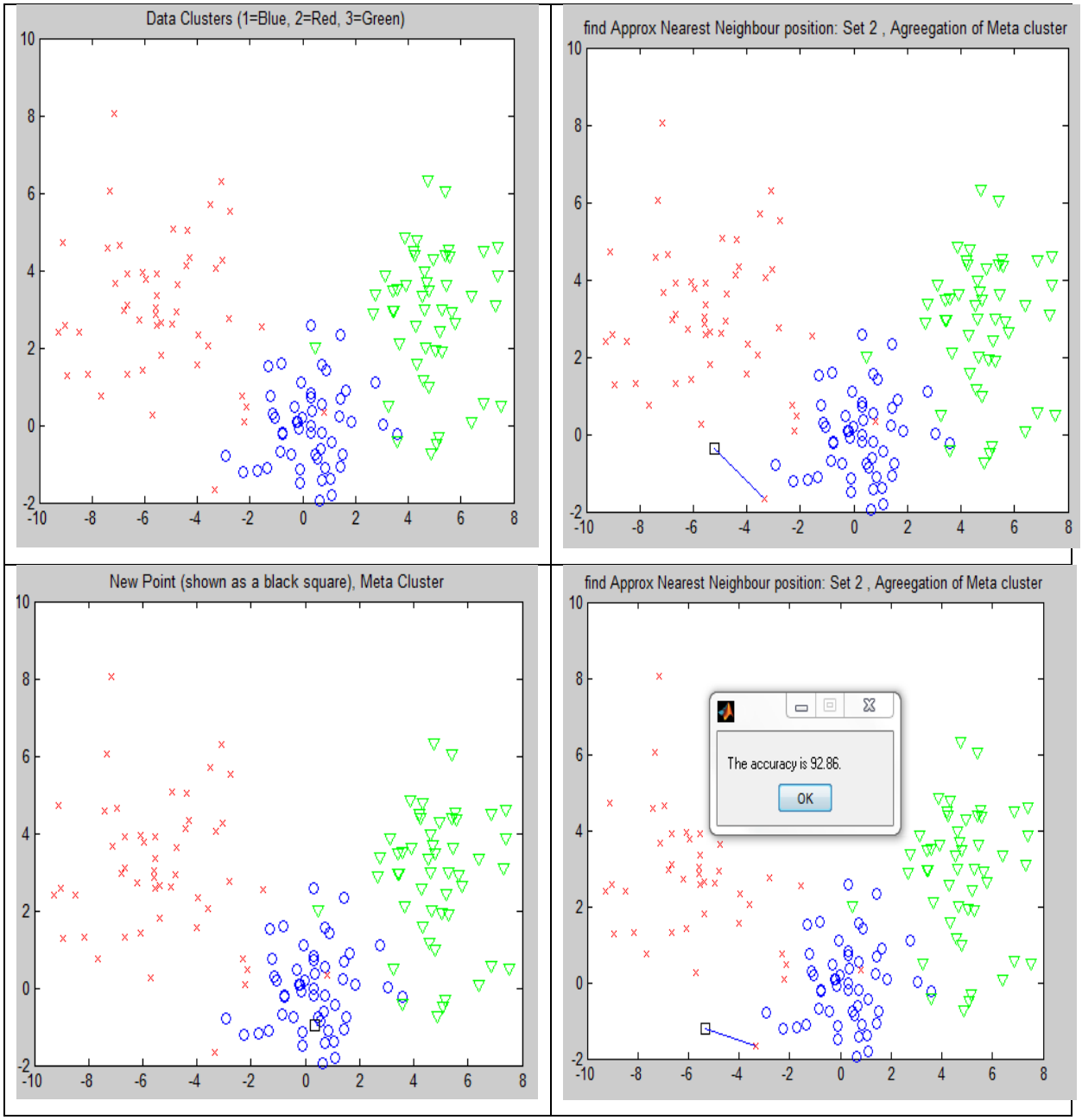


**Fig 4.6: Normalization is applied for central point selection and cluster data**

As shown in the figure 4.6, the dataset is loaded and on the loaded dataset number of iterations are selected, the HMAC is applied and Meta clustering with k-mean and density based clustering is applied. After clustering the data normalization point is selected for the number of time for clustering the data. 3NF polynomial normalization is applied on the data sets to cluster the data elements. This 3NF polynomial normalization produces more accurate results of clustering. Normalization point moves to cluster the data. Time to time it moves on different location to cluster the data elements. With the help of normalization point unique elements are discovered. Central points are also discovered with the help of normalization point and sigma functions also help to discover the central points to cluster the data elements. This central point helps to cluster the data elements into number of clusters.

As shown in the figure 4.7, the dataset is loaded and on the loaded dataset number of iterations are selected, the HMAC is applied and Meta clustering with K-mean and density based clustering is applied. After clustering the data normalization point is selected for the number of time for clustering the data. Position of normalization point is change with respect to time. In this diagram normalization point is shown with the help of Black Square. Black square when moves means normalization point is moving to find the most relevant point to cluster the data elements that is present in data sets. Meta clustering is used to cluster the base level clusterings. Firstly base level clusterings are produced by applying K-mean clustering algorithm which is based on random partitions of the data elements. Data sets are clustered using K-mean clustering and then applying Meta clustering to group the similar clusters into a single group. After applying these two clusterings normalization is performed to find the unique points and cluster data accurately. Normalization point helps to cluster the data efficiently. Normalization point plays a very important role to cluster the similar data elements into single cluster and remove the different data elements as noise. KL-divergence with density based clustering and enhanced Meta clustering with K-mean clustering is given below in diagram. KL-Divergence with density based clustering can't handle complex data efficiently. Enhanced Meta clustering cluster data efficiently. Meta clustering produces 57.14% accuracy, but on the other hand enhanced Meta clustering produces 92.86% accuracy as in diagram 4.7 given below. There is greater difference in accuracy enhanced Meta clustering produce more accurate results than that of Meta clustering. It is not good in handling uncertain data items.

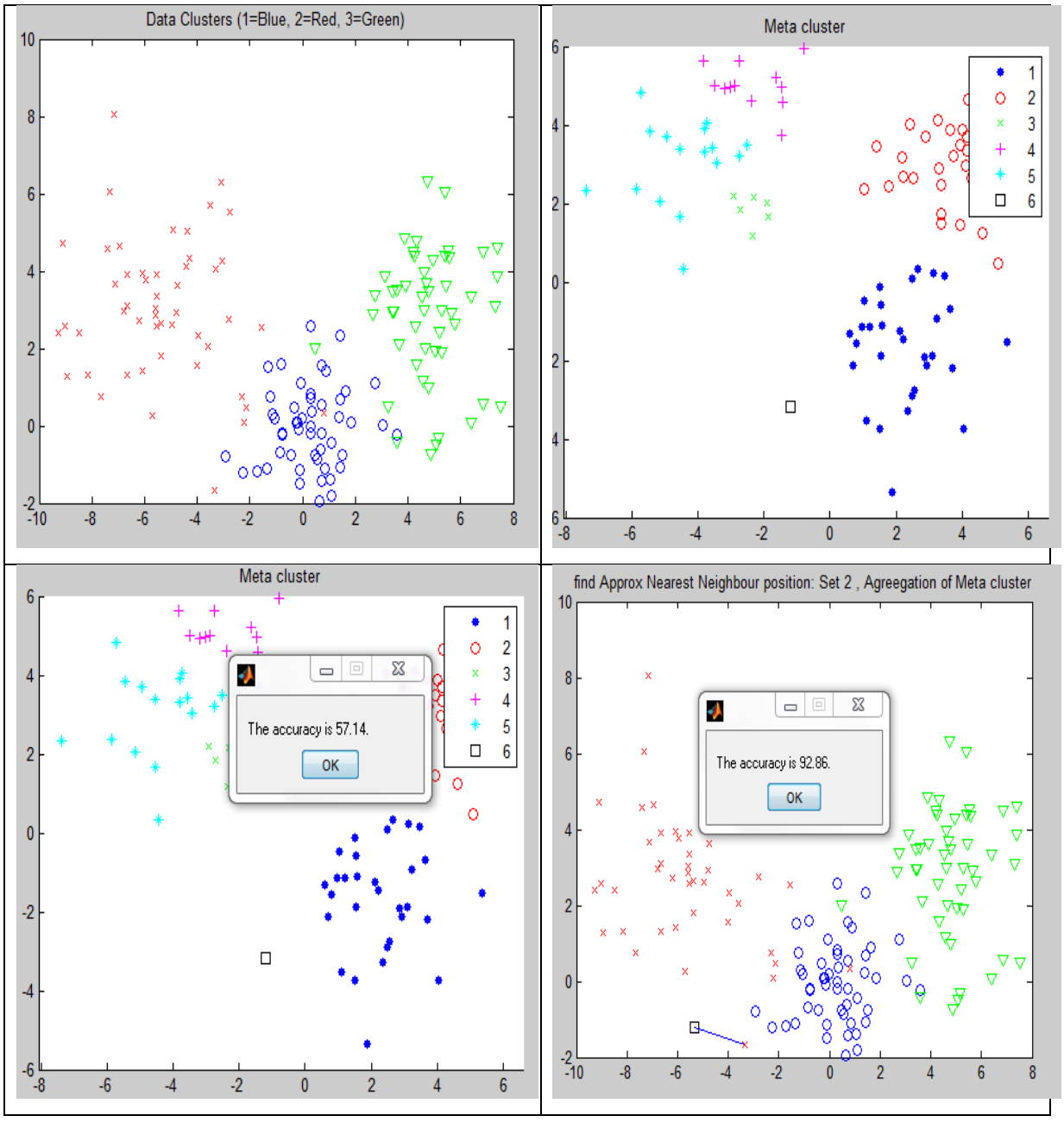


**Fig 4.7: Clustered Data**

As illustrated in figure 4.8, the normalization point is selected number of time to improve the cluster quality. In this figure final output of the clustered dataset is shown on 2D plane. Both techniques are as given in figure 4.7 as shown above. Both of them give different accuracy and different timing. KL-Divergence cluster data elements in different way by using density based clustering as shown above in figure 4.7. Both of them give different accuracy as shown above in figure.

Clustering using enhanced meta-clustering give good results in term of accuracy and in term of quality. Old k-mean clustering and density based clustering algorithm not give the efficient and quality results while KL-Divergence method with K-mean gives quality data. KL-Divergence measures the similarity between the data items that are uncertain in nature. After that density based or partitioning clustering method applied. It gives quality results but not applied on complex data sets. Enhanced meta-clustering produce good results in term of accuracy and also processing time would decrease as shown in graph given below. Accuracy values are given in graph as shown in fig 4.8.

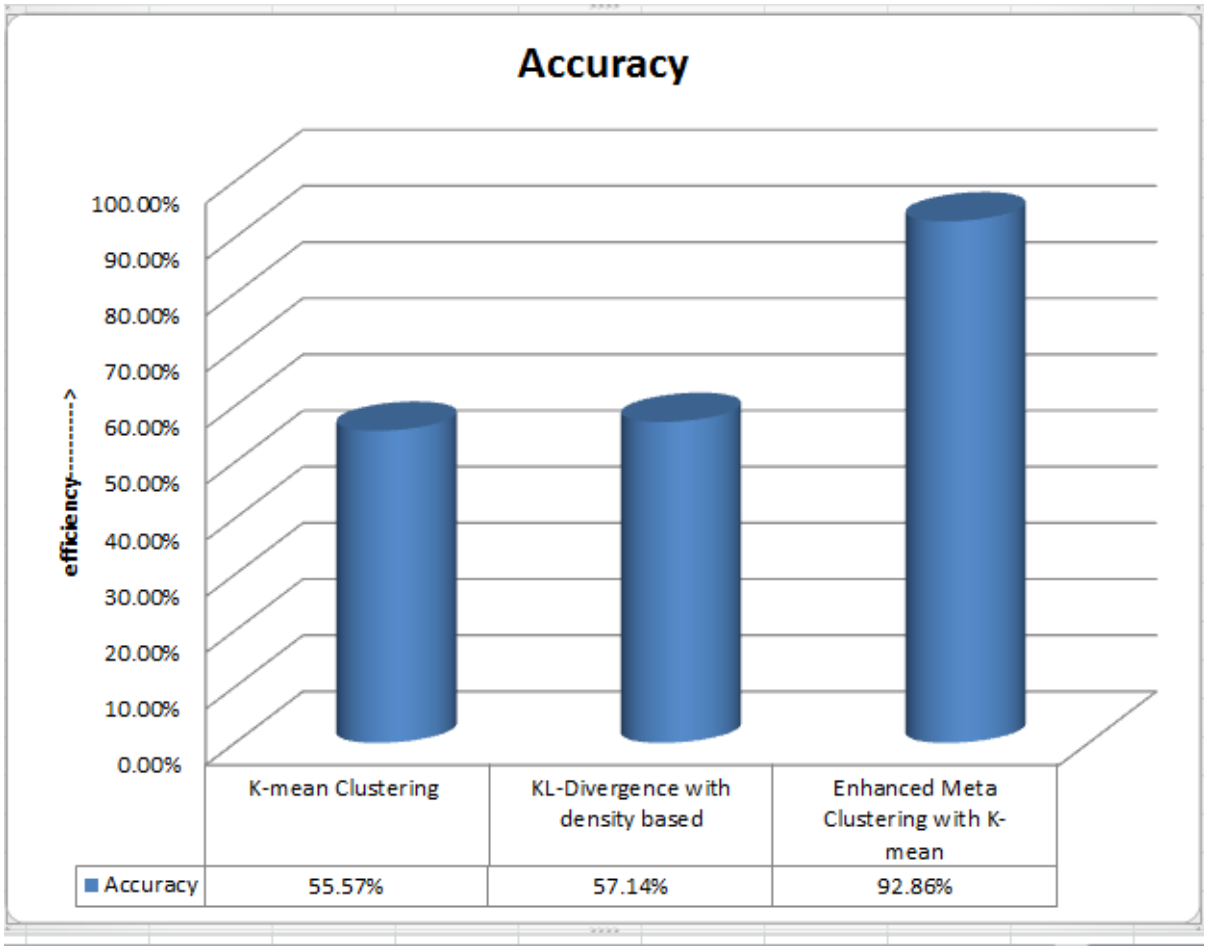


**Fig.4.8.Accuracy comparison**

Given graph illustrates the accuracy comparison in terms of quality of clusters. As shown above K-mean clustering when applied on uncertain data it would not produce good results of cluster quality. Uncertain data are not easy to handle. It is difficult to handle uncertain data

using K-mean. KL-divergence with k-mean or density based clustering would produce better results than K-mean clustering but enhanced meta-clustering with K-mean produces more accurate results. It handles uncertain data very easily. KL-divergence with density based clustering gives 57.14% accuracy and enhanced meta-clustering gives 92.86% accuracy as shown in graph. Comparison of these techniques is given above. Enhanced meta-clustering is very good in handling uncertain data items.

**Time comparison: -** Time taken by both of the cases i.e. KL-Divergence with density based clustering and Meta-clustering with K-mean is not same in all the time when executes. Dynamic time is there when both of the clustering execution done. KL-divergence with density based clustering always takes more time than that of enhanced meta-clustering.

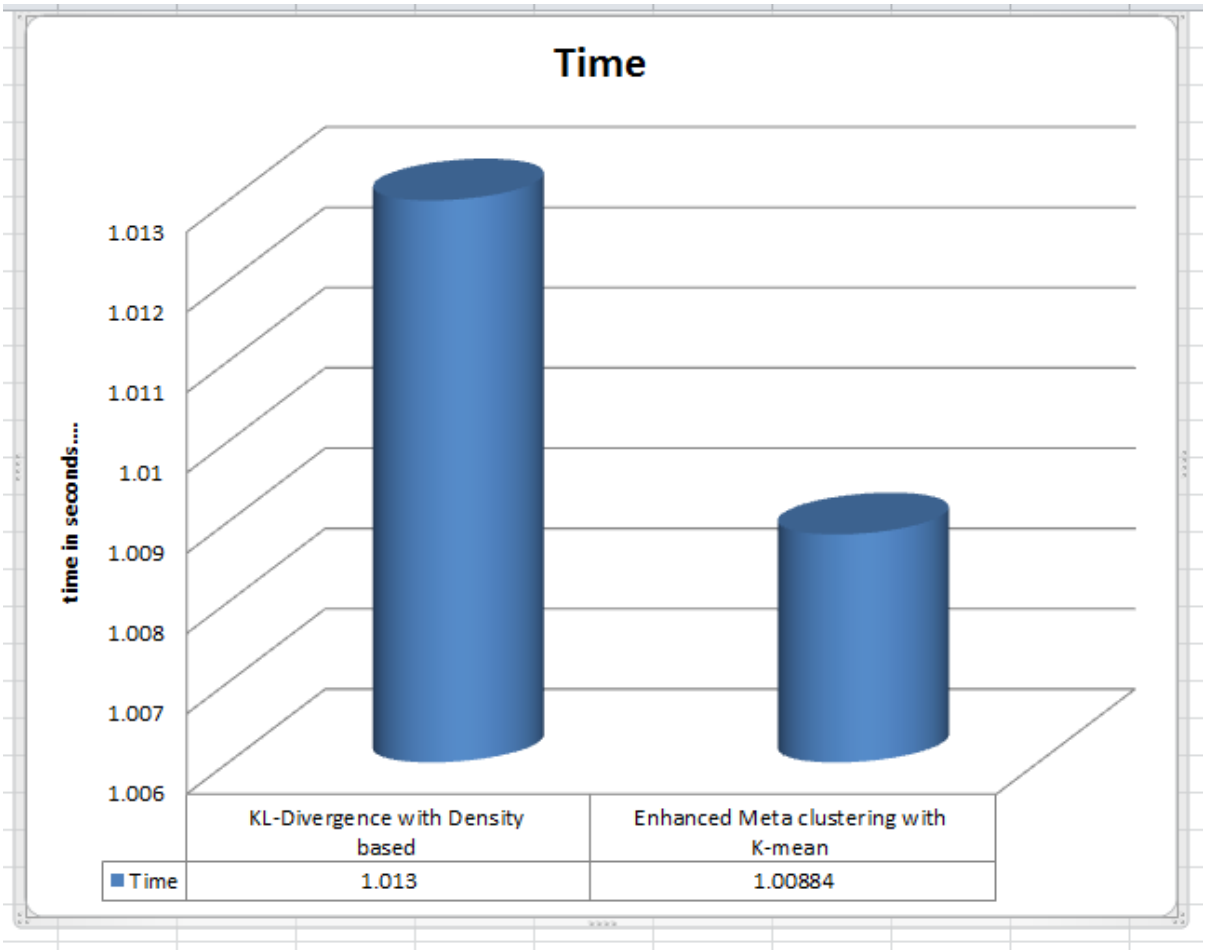


**Fig4.9 Time comparison of KL-divergence and Meta-clustering**

Enhanced meta-clustering takes lesser time to handle uncertain data items than KL-divergence divergence with density based clustering. KL-divergence is more complex and time consuming there is no large difference of timing in both of the clustering techniques. A small difference of timing is there. When it executed firstly KL-divergence with density based clustering takes 1.014 times and Meta-clustering with k-mean takes 1.0133 to cluster uncertain data. Timing is not always same in all the cases when executed. Dynamic time is there. Time varies with execution.

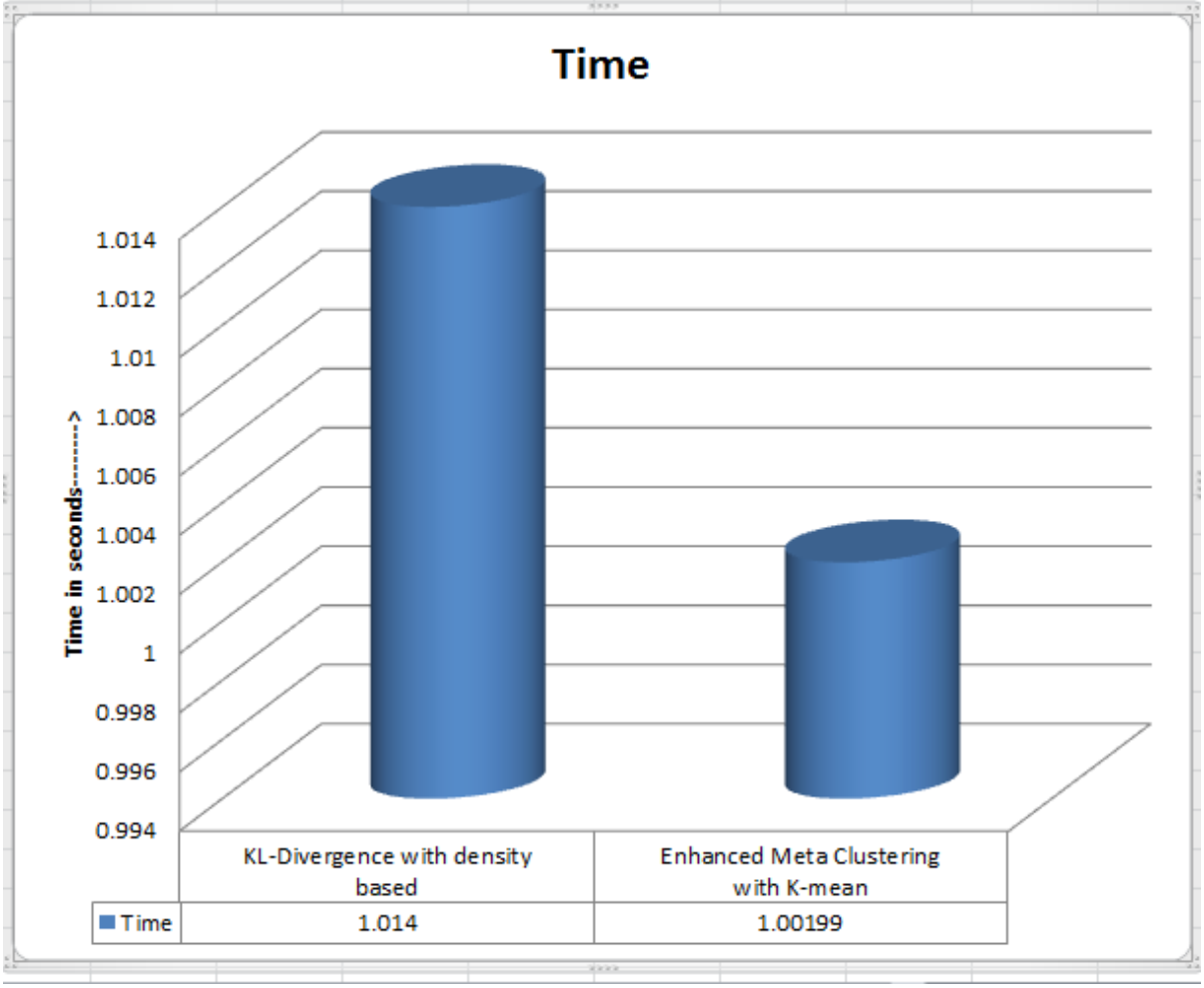


**Fig4.10: Time complexity in KL-divergence and Meta clustering with k-mean**

Time varies according to execution of implemented clustering program. Data takes different time every time when clustered. In this graph KL-Divergence with density based clustering takes more time to cluster data and also accuracy of this algorithm is also lesser than that of

enhanced Meta Clustering. Enhanced Meta-clustering is more accurate its accuracy is 92.84% and takes less time than that of KL-Divergence with density based clustering.

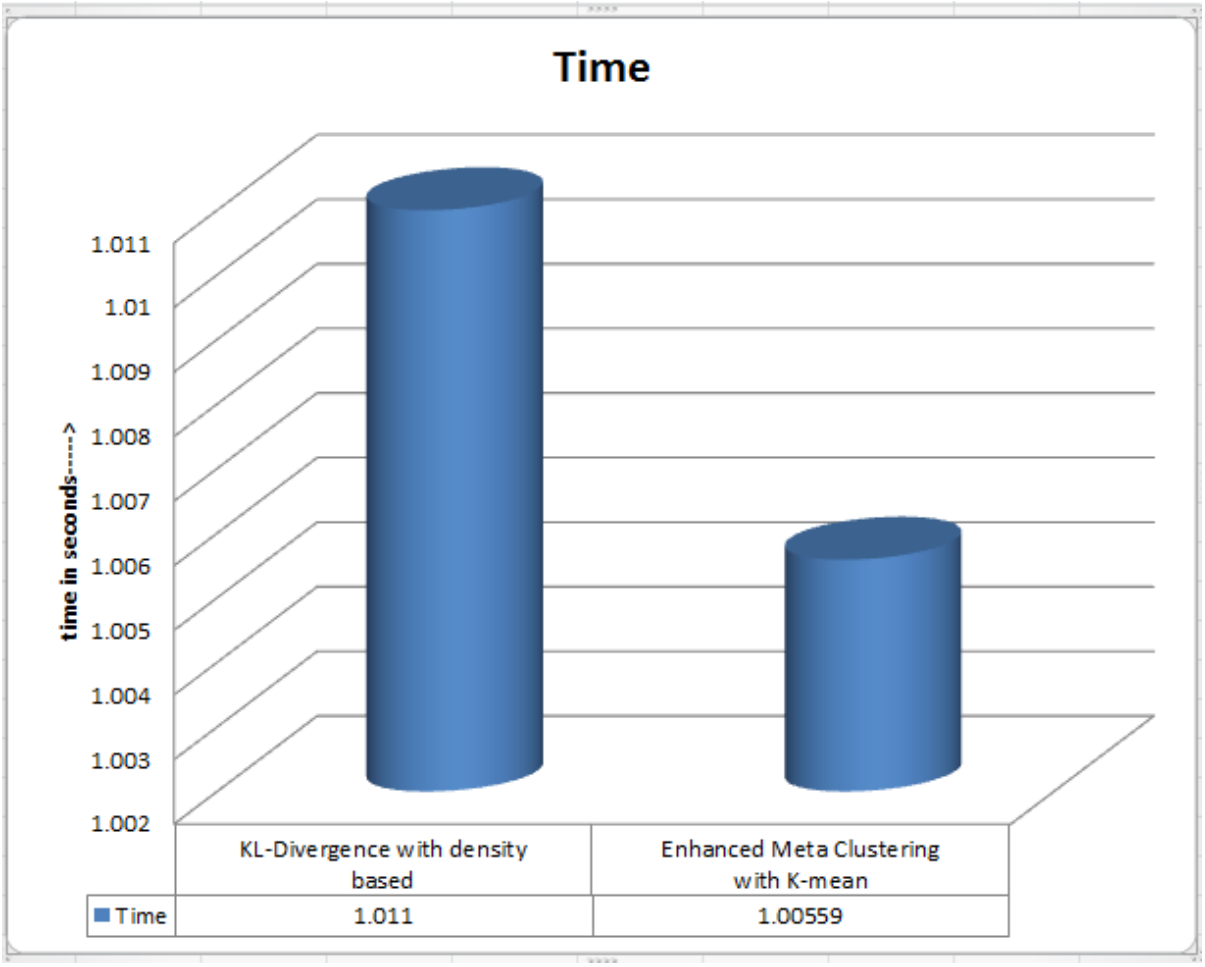


**Fig4.11.Dynamic time comparison of KL-Divergence with density based clustering and Enhanced Meta clustering with K-mean**

This is one more output of timing comparison of KL-Divergence with density based clustering and enhanced Meta-Clustering with K-mean clustering. K-mean clustering partitioned the data element in different way every time when cluster the data and take different time every execution of clustering. That's why clustering time is dynamic in nature. It changes with each different execution. In this case KL-divergence takes 1.011s and enhanced Meta clustering takes 1.00559s time. This time changes in every execution. This change in time is just because of K-mean clustering because it takes different partitions every

time for cluster data elements on random basis. Different readings of the time can be taken and new graphs of that reading are there as shown below in figure.

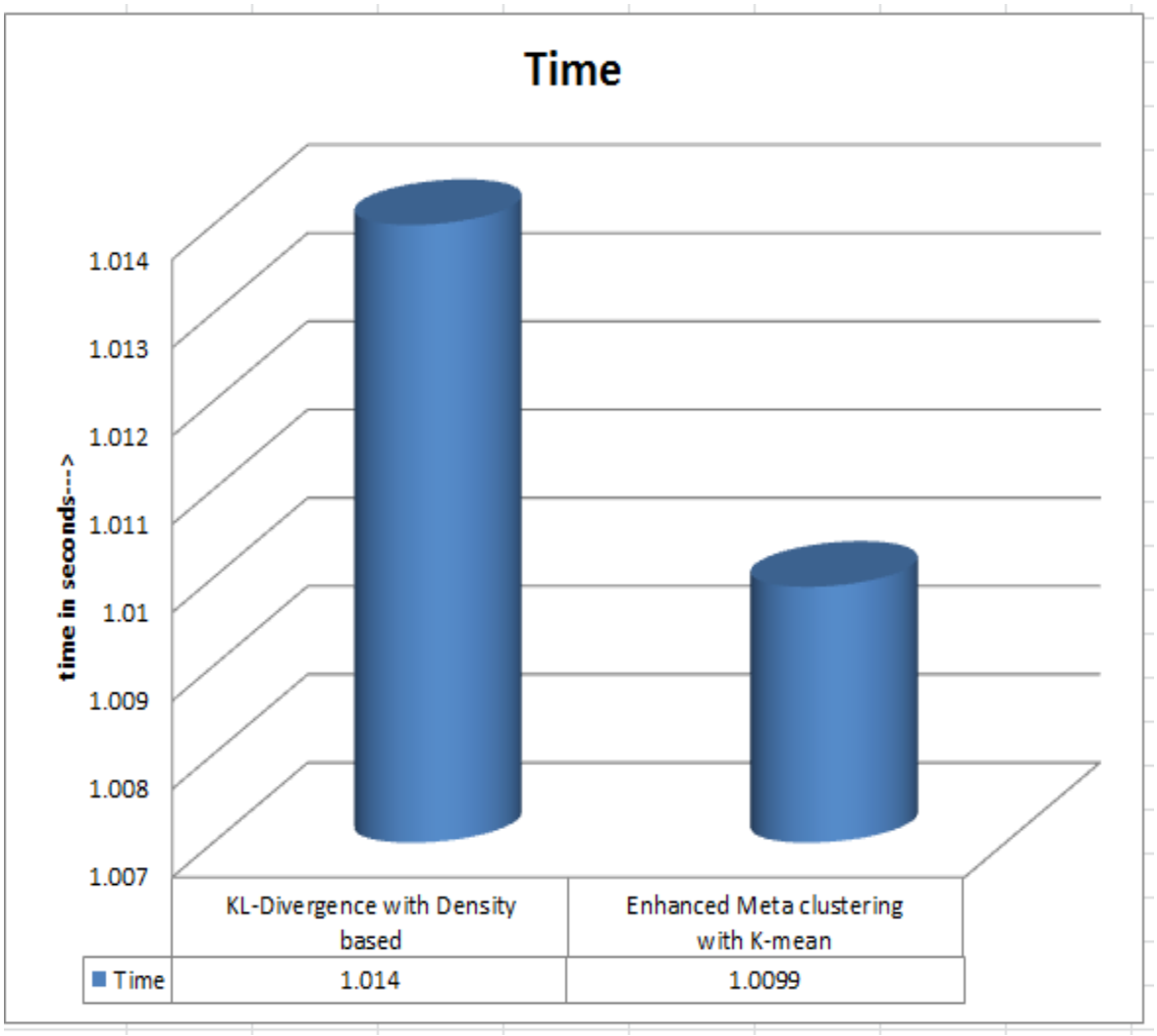


**Fig. 4.12 New times for both KL-Divergence and Enhanced Meta clustering**

In this given graph new time for both KL-divergences with density based clustering and enhanced Meta clustering with K-mean clustering. Old technique i.e. KL-divergence with density based clustering takes 1.014 seconds to cluster the data elements into groups or clusters. Purposed new technique i.e. enhanced Meta clustering with K-mean clustering to cluster the data elements takes 1.0099 seconds. Hence it concluded that KL-Divergence takes more time to cluster the data elements. Time varies with different executions; it changes at every execution of clustering just because of K-mean clustering. K-mean clustering

technique cluster data every time by selecting different data elements for initialization. N number of readings can be taken by executing the implemented technique again and again different time would produce every execution as shown below in graph new readings are there. K-mean clustering is applied after loading the data set and adjusting the elements into a single hierarchy. K-mean Clustering technique would cluster the data elements into single clusters that are similar to each other. K-mean clustering selects initial partition on the random bases. Each time different elements are selected. Hence time would decrease and increase at every execution as shown below in new graph.

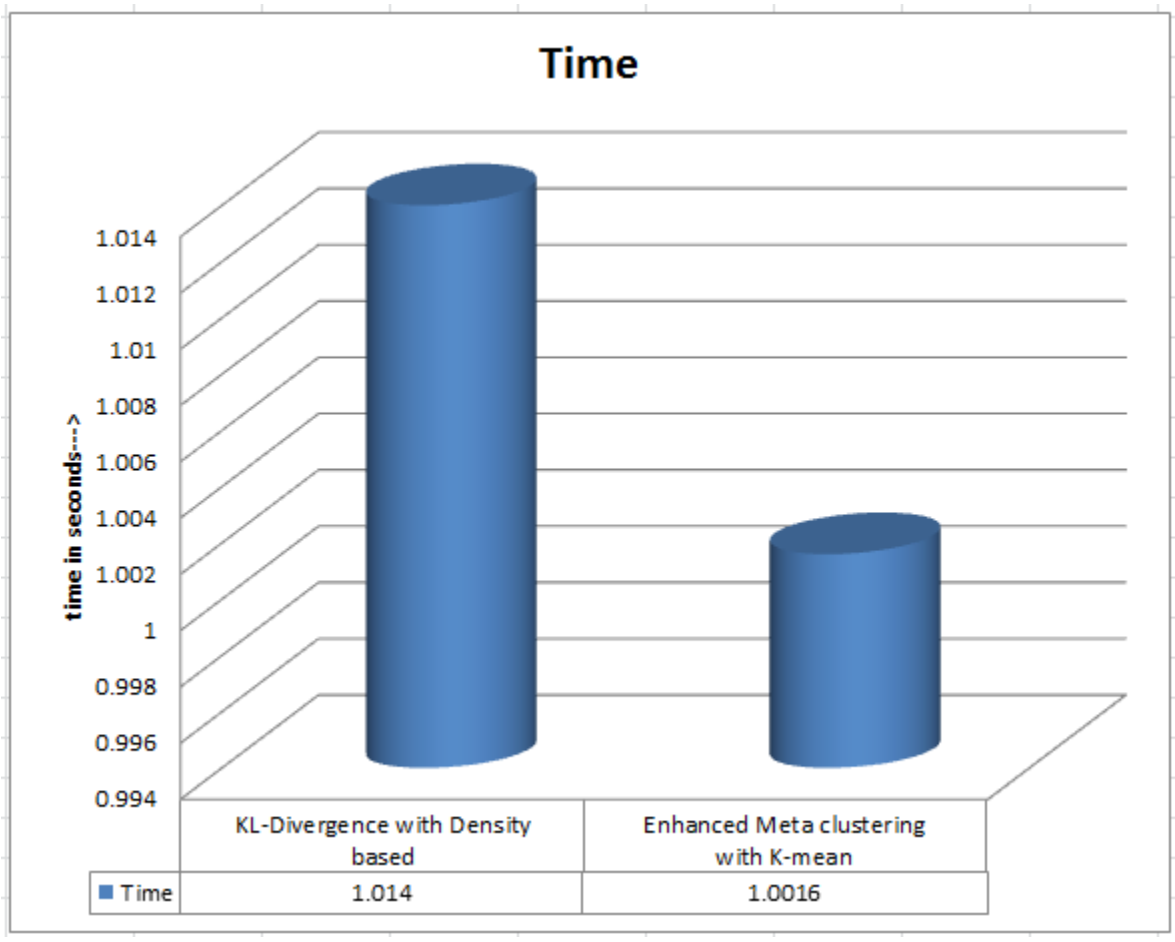


**Fig. 4.13 New times for both KL-Divergence and Enhanced Meta clustering**

In this case KL-Divergence with density based clustering takes 1.014 seconds to cluster the data elements that are similar to each other into a single cluster. Enhanced Meta clustering takes 1.0016 seconds to cluster the data elements. As discussed above time is dynamic in nature in this clustering because K-mean clustering is used to cluster the data elements.

# CHAPTER-5
# CONCLUSION AND FUTURE SCOPE

## CONCLUSION

Data is increasing day by day in volume, variety and velocity. Data has basically three features volume i.e. data is very big in size when collected from various resources. Second feature of data is velocity means data is moving in nature. Third feature of data is variety i.e. data is varying in nature it varies day by day. Actually small amount of data is usable not all data collected is usable. Simply data volumes are very large and are difficult to manage these data becomes complicated structures and becomes more complex. Data mining is used to take only usable data from these volumes of data. Data mining is just because to understand the nature of data and to handle data. Today's competitive world data mining is very important i.e. extract beneficial knowledge hidden in large volumes of data. Computer based methodology and new approaches are used to realize knowledge from large volume data is called as data mining. Data gets clustered using clustering techniques. Data is grouped into clusters, a cluster contains more similarity of data items with each other and outside the cluster data items are dissimilar to each other. That there is high intra cluster similarity and high inter cluster dissimilarity. There is countless clustering method such as partitioning method, hierarchical method, density based method, and grid based method. Uncertain data items are clustered using density based clustering but it can't handle the uncertain data items easily because data are geometrically indistinguishable and takes more efforts to cluster such data. Density based clustering comprise mainly two algorithms i.e. DBSCAN and OPTICS these are basic used algorithm of density based clustering. For handling uncertain data which is geometrically indistinguishable KL-Divergence and density based clustering combination is used which is good for handling uncertain data. Further enhancement in algorithm is required because there are some accuracy problems. Purposed work enhancing the clustering algorithm and improving the cluster quality. Old technique is less efficient to handle uncertain data items. Enhanced Meta clustering is used to handle uncertain data items. It would produce more accurate results in handling uncertain data items. Results and discussion describes about the accuracy of both the algorithms. Enhanced Meta clustering produce

92.86% accuracy it would handle uncertain data items more accurately and efficiently. But on the other hand Meta clustering would not give accurate results. Its accuracy is only 57.14%; there is a lot difference in time of execution of both the techniques as shown in graphs. Dynamic time of execution is there. Always it takes different time when executed. This variation in time is just because of using K-mean clustering algorithm to cluster the data elements.

**FUTURE SCOPE: -** In this work purposed method to handle uncertain data is very effective in nature. It produces the accurate results. Means its output shows efficiency in it. Data clustered in this technique take less time than that of KL-Divergence technique. Its accuracy is 92.84% while old technique shows only 57.14% accuracy. KL-Divergence technique does not produce effective results in term of accuracy. Time taken by both the techniques is dynamic in nature. It will not remain same all the times when executing the clustering implementation of enhanced Meta clustering and KL-divergence with density based or K-mean clustering. Time is dynamic this is the limitation of the new technique. Time should be stable or say static it should not vary when execution of the implemented program of clustering. It is dynamic in nature because K-mean clustering is used. K-mean clustering select randomly data items initially for cluster data objects hence take different time for clustering each execution. K data items are selected firstly according to K-mean clustering. This selection is on the basis of random selection. This is the basic reason for dynamic time in cluster the data values.

# CHAPTER-6
# LIST OF REFERENCES


[1] Ashish Patel (2014), "Density Based Clustering Based on Probability Distribution for Uncertain data

[2]Ahmed Rafea and Nada A. Mostafa (2014), "Topic Extraction in Social Media"*978-1-4673-6404- ©IEEE.*

[3] Alvaro Garcia-Piquer and Albert Fornells (2014) "Large-Scale Experimental Evaluation of Cluster Representations for Multi-objective Evolutionary Clustering" *IEEE Transaction on Evolutionary Computation, FEBRUARY*

[4] Alessia Albanese (2014) Member of IEEE, Sankar K. Pal, "Rough Sets, Kernel Set, and Spatiotemporal Outlier Detection*"IEEE Transaction on Knowledge and Data Engineering, 1, JANUARY.*

[5] Aliya Edathadathil1 (2014), "A Modified *K*-Medoid Method to Cluster Uncertain Data Based on Probability Distribution Similarity", *International Journal Of Computer Science Engineering And 7 July,.*

[6] Anirban Mukhopadhyay (2014) *Senior Member of IEEE,* Ujjwal Maulik, *Senior Member in IEEE* "Survey of Multi-objective Evolutionary Algorithms for Data Mining: Part II" *IEEE Transaction on Evolutionary Computation,.*

[7] Ana L.N. Fredand Anil K. Jain, "Combining Multiple Clustering Using Evidence Accumulation."

[8] ArnostKomarek and LenkaKomarkova (2013), "Clustering for Multivariate Continuous and Discrete Longitudinal Data" *The Analysis of Applied Statistics, Vol. 7, No. 1, 177–200.*

[9] ARISTIDES GIONIS and HEIKKI MANNILA, "Clustering Aggregation" ACM Journal Name.

[10] Bin Jiang and Jian Pei (2011), "Clustering Uncertain Data Based on Probability Distribution Similarity", *Digital Object analysis 10.1109/TKDE.2011.221 1041-4347/11/$26.00 © IEEE.*

[11] Bo Liu, Yanshan Xiao (2011),"Uncertain One-Class Learning and Concept Summarization Learning on Uncertain Data Streams" *IEEE Business on Information and Data Engineering, JANUARY* .

[12] Cane Wing-Ki Leung and Stephen Chi-fai Chan (2011), "A probabilistic rating inference framework for mining user preferences from reviews" *Springer@ Science and Business Media, LLC.*

[13] C.Deepika (2014), "An Efficient Uncertain Data Point Clustering Based On Probability–Maximization Algorithm", *An ISO 3297: 2007 Certified Organization Vol. 2, Issue 8, August.*

[14] D. M. PAdulkar, V. Z. Attar (2012), "Uncertain Numerical Data Clustering with VORONOI Diagram and R-Tree with Ensemble SVM", *International Journal of Computer Science and Information Technologies, Vol. 3 (3), 2012, 4549-4552.*

[15] Dimitrios Mavroeidis and Elena Marchiori (2014), "Feature selection for k-means clustering stability: theoretical analysis and an algorithm*", Data Min Knowl Disc 28:918–960 DOI 10.1007/s10618-013-0320-3 springer.*

[16] Francesco Gullo (2012), "Uncertain Centroid based Partitioned Clustering of Uncertain Data"*August 27th 31st, Istanbul, Turkey. Proceedings of the VLDB Endowment, Vol. 5, No. 7.*

[17]Fatih Dikbas and MahmutFirat (2013), "Defining Homogeneous Regions for Streamflow Processes in Turkey Using a K-Means Clustering Method" *King Fahd University of Petroleum and Minerals.*

[18] Feng CaoMartin Ester (2013), "Density-Based Clustering over an Evolving Data Stream with Noise."

[19] F.U. SIDDIQUI and N.A. MAT ISA (2013), "Optimized K-means (OKM) clustering algorithm for image segmentation"*OPTO−ELECTRONICS REVIEW, 216–225 DOI: 10.2478/s11772−012−0028−8.*

[20] G. W. Ma & Z. H. Xu & W. Zhang & S. C. Li (2014), "An enriched K-means clustering method for grouping fractures with meliorated initial centers*", Springer Saudi Society for Geosciences.*

[21] Graham Cormode (2008), "Approximation Algorithms for Clustering Uncertain Data*", Vancouver, BC, Canada.*

[22] Jean Christoph Jung, Carsten Lutz University Bermen, Germany, "Ontology-Based Access to Probabilistic Data".

[23] Lipika Dey and Sk. MirajulHaque (2009), "Opinion mining from noisy text data" *Published online: 21 August, Springer-Verlag.*

[24]Markus M. Breunig, Hans-Peter Kriegel (2000), "LOF: Identifying Density-Based Local Outliers"*MOD 2000, Dallas, TX USA © ACM.*

[25] Michael Chau1 and Reynold Cheng (2005)., "Uncertain Data Mining: An Example in Clustering Location Data", *Workshop on the Sciences of the Artificial, Hualien, Taiwan* .

[26] Patrick Glenisson and Janick Mathys, "Meta-Clustering of Gene Expression Data and Literature-based Information"*SIGKDD Explorations.Volume 5.*

[27] Reshma MR (2013), "Handling Uncertainty and Clustering Uncertain Data based on KL-Divergence Technique"*IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555Vol. 3, No.5, and October* 365.

[28] Rich Caruana and Mohamed Elhawary, Nam Nguyen, Casey Smith, "Meta Clustering", *Cornell University, Ithaca, New York 14853.*

[29] Samir N. Ajani and Prof. MangeshWanjari (2013), "Clustering of Uncertain Data Objects using Improved K-means Algorithm",*International Journal of Advanced Research in Computer Science and Software Engineering,Volume 3, Issue 5, May.*

[30] Tina Eliassi-Rad Terence Critchlow (2007), "Multivariate Clustering of Large-Scale Scientific Simulation Data", *In Proceedings of SIGMOD Record, ACM Press 28(4):49-57, 2007.*

[31] Tayfun DOGDAS and Selim AKYOKUS (2013),"Document Clustering using GIS Visualizing and EM Clustering Method" *©2013 IEEE.*

[32] T.HITENDRA SARMA and P.VISWANATH (2013),"Single pass kernel *k*-means clustering method", *Indian Academy of Sciences.*

[33] Weifang SHI and Weihua ZENG (2013),"Application of k-means clustering to environmental risk Zoning of the chemical industrial area*", Higher Education Press and Springer-Verlag Berlin Heidelberg.*

[34]Xiaoli Cui, Pingfei Zhu and Xin Yang (2014), "Optimized big data K-means clustering using Map-Reduce", *Springer* .

[35] Yi Ma (2007), "Segmentation of Multivariate Mixed Data via Lossy Data Coding and Compression".

[36] Yi Zhang and Tao Li (2013), "Consensus Clustering + Meta Clustering= Multiple Consensus Clustering".

[37] Zhaohong "A Survey on Soft Subspace Clustering."

[38] Z. Volkovich and D. Toledano-Kitai (2012), "Self-learning *K*-means clustering: a global optimization Approach", *7 February Springer Science and Business Media, LLC.*

# CHAPTER-7
# APPENDIX

**LIST OF ABBRIVIATIONS**

| | |
|---|---|
| D. | Set of data containing n objects. |
| DBSCAN | Density based clustering algorithm |
| K | Number of clusters |
| KDD | Knowledge Discovery Data |
| OPTICS | ordering points to identify the clustering Structure. |
| CLARANS | Clustering large applications based on randomized research. |
| OBDA | Ontology based data access |
| HSC | Hard subspace clustering |
| SSC | .Soft subspace clustering. |
| ISSC | Independent soft subspace clustering. |
| CSSC | Conventional Soft subspace clustering |
| XSSC | Extended Soft Subspace clustering |
| N | number of objects in data set |
| KL-Div | Kulback Libeler-Divergence |
| EM | Expectation maximization |
| GIS | geographical information system |
| 2D | Two Dimensional Planes |
| LOF | Local Outlier factor |

PDBMS Probabilistic Database management system

FCM Fuzzy C-Mean

3D Three Dimensional plane

HC Hierarchical clustering

GC Grid Clustering

PC Partitioning Clustering

PM Probability maximization

CF Collaborative Filtering

## LIST OF PAPERS PUBLISHED

**Enhancement in clustering to find the cluster center and improve the cluster quality**

**Abstract: -** In daily life more much amount of data is produced. Sometimes data created is uncertain and is difficult to handle. By using clustering uncertain data can be handled. Determining cluster center or say guesstimate of cluster center issues from these dissemination are expected. Problem is cracked by non-linear minimum square optimization compliant the cluster center and cluster size. Clustering has many applications in many domains Where clustering is used in various fields of our real life. It is a development which is given here and described few on it. Also is used in various fields are unknown on it. Finding cluster center and cluster sizes are very useful in many areas. Purposed work would describe the enhancement of clustering to define the cluster center to increase cluster quality. Suggested technique is very useful for practical applications and theoretical reflection for clustering problems